NASA-TM-112630

REPORT CONCERNING SPACE
DATA SYSTEMS STANDARDS

Reference Model for an Open Archival Information System

CCSDS 650.0-W-1

WHITE BOOK

April 10, 1997

Dear Reader:

The following version of the OAIS Reference Model is designated as the first CCSDS White Book designated as CCSDS 650.0-W-1. This book is a significant improvement over previous versions and is considered ready for exposure to external reviewers. There are several known deficiencies:

- Đ the glossary is still incomplete
- Đ section 3.1.8 context diagrams need another pass and additional text
- Đ section 4. on migration is new and may need enhancement from the scenarios from Annex B
- Đ section 5 on classification is new
- Đ section 6, the illustrative scenario, needs to be extended to include migration and more on access
- Đ scenarios in Annex A need to be updated to include the information model framework
- Đ Annex B will probably be merged into section 4
- Đ Annex C needs to be updated or eliminate
- Đ Annex D on a guide to OMT needs to be supplied

In addition, there are two major technical issues for which white papers are being prepared. These issues are Ôdifferentiating leaf level AICs from higher level AICsÕ and Ôshowing that all preservation descriptions also require representation informationÕ. These papers will be available prior to the international workshop.

This will be the last version prior to the international workshop. All US and international experts are requested to provide comments prior to May 3, 1997. Comments may be sent to:

Lou Reich
louis.i.reich@gsfc.nasa.gov

or

Don Sawyer
donald.sawyer@gsfc.nasa.gov

A new version of this book is anticipated for release prior to July 1, 1997.

CCSDS REPORT- REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM

AUTHORITY

Issue:	White Book, Issue 1
Date:	April 10, 1997
Location:	Oberpfaffenhofen, Germany

This document, when it has been approved for publication by the Management Council of the Consultative Committee for Space Data Systems (CCSDS), will reflect the consensus of technical panel experts from CCSDS Member Agencies. The procedure for review and authorization of CCSDS Reports is detailed in reference [1].

This document is published and maintained by:

CCSDS Secretariat
Program Integration Division (Code MG)
National Aeronautics and Space Administration
Washington, DC 20546, USA

FOREWORD

This document is a technical Report for use in developing a consensus on what is required to operate a permanent, or indefinite long-term, archive of digital information. It may be useful as a starting point for a similar document addressing the indefinite long-term

preservation of non-digital information.

This Report establishes a common framework of terms and concepts which comprise an Open Archival Information System (OAIS). It allows existing and future archives to be more meaningfully compared and contrasted. It provides a basis for further standardization within an archival context and it should promote greater vendor awareness of, and support of, archival requirements.

Through the process of normal evolution, it is expected that expansion, deletion, or modification to this document may occur. This Report is therefore subject to CCSDS document management and change control procedures, which are defined in Reference [1].

DOCUMENT CONTROL

Document	Title	Date	Status and Substantive Changes
C C S D S 650.0-W-1	Report Concerning Space Data Systems Standards: Reference Model for an Open Archival Information System (OAIS)	A p r i l 1997	Original Issue

CONTENTS

CCSDS REPORT- REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM

2.4.6 MAKES THE INFORMATION AVAILABLE

CONTENTS (CONTINUED)

Section Page

CONTENTS (continued)

Section Page

ANNEX C: COMPATABILITY WITH OTHER STANDARDS

ANNEX D: BRIEF GUIDE TO THE OMT

1 INTRODUCTION

1.1 PURPOSE AND SCOPE

The purpose of this document is to define the ISO Reference Model for an Open Archival Information System (OAIS). This reference model:

- Ð provides a framework for the understanding of archival concepts needed for permanent, or indefinite long-term, digital information preservation. 'Long-term' is long enough to be concerned with the impacts of changing technologies including support for new media and data formats. Shorter term archives may find much of the framework useful as well.

- Ð provides a framework that facilitates the description and comparison of the architecture's and operations of existing and future information-preserving archives, and it provides a basis for comparing the data modelling and transformation aspects of the digital information preserved and managed by these archives.

- Ð provides a starting point that may be expanded by other efforts to cover long-term preservation of information that is NOT in digital form (e.g., physical media, physical samples);

- Ð expands consensus on the requirements for long-term digital information preservation and should lead to a larger market which vendors can support;

- Ð guides the production of OAIS related standards.

The ISO Reference Model for an Open Archival Information System defines a minimum set of responsibilities for the recognition of an OAIS archive. This allows an OAIS

archive to be distinguished from other uses of the term 'archive.'

The Reference Model addresses a full range of archival information preservation functions including ingest, storage, access, and dissemination. It also addresses the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. It identifies both internal and external interfaces to the archive functions, and it identifies a number of high level services at these interfaces.. A first step toward establishing the ability to characterize various types of archives, and their quality, is also addressed.

1.2 APPLICABILITY

The OAIS model in this document is applicable to organizations with the responsibility of making information available permanently or for the indefinite long-term. This model is also of interest to those organizations and individuals who create information that may need indefinite long-term preservation and those that may need to acquire information from such archives. This model may also be useful for those organizations developing shorter-term archives, or repositories, especially when taking into consideration the rapid pace of technology and the likelihood that many repositories thought of as temporary will in fact find that some or much of their holdings will need the same type of attention as that given by permanent archives.

Standards developers are expected to use this model as a basis for further standardization and therefore provide an extension of what is meant by "operating an OAIS archive". A large number of related standards are possible. A road-map for such development is briefly addressed in section 1.4

This Reference Model does not specify an implementation. Actual implementations may group or break out functionality differently.

1.3 RATIONALE

A tremendous growth in computational power, and in networking bandwidth and connectivity, have resulted in an explosion of organizations who are making information available in electronic forms. Transactions among all types of organizations are being conducted using electronic forms that are taking the place of more traditional forms such as paper.

Preserving information in electronic forms is much more difficult than for forms such as paper and film. This is not only a problem for traditional archives, but for many

organizations that have never thought of themselves as performing an archival function. It is expected that this reference model, by establishing minimum requirements for an OAIS archive along with a set of archival concepts, will provide a common framework from which to view archival challenges, particularly as they relate to digital information. This should enable more organizations to understand the issues and to take proper steps to ensure long term information preservation. It should also provide a basis for more standardization and therefore a larger market that vendors can support in meeting archival requirements.

1.4 ROAD-MAP FOR DEVELOPMENT OF RELATED STANDARDS

This Reference Model serves to identify areas suitable for the development of OAIS related standards. Some of these standards may be developed by Panel 2 of the CCSDS (sub-committee of ISO); others may be developed by other standardization bodies. However, any such work undertaken by other bodies should be coordinated with CCSDS Panel 2 in order to minimize incompatibilities and efforts. Areas for potential OAIS related standards include:

- standard(s) for the interfaces between OAIS type archives.
- standard(s) for the submission (ingest) of digital data sources to the archive.
- standard(s) for the delivery of digital sources from the archive to data users.
- standard(s) for the submission of digital metadata about digital or non-digital data sources to the archive. Here, it can be envisaged that different disciplines might require different metadata standards.
- protocol standard(s) to search and retrieve metadata information about digital or non-digital data sources. Here, the adoption of profiles of the Z39.50 search and retrieval protocol may be considered. The Catalogue Interoperability Protocol (CIP) and the Digital Collection (DC) profile may be suitable candidates.
- standard(s) for media access allowing replacement of media management

systems without having to re-write the media

- Ð standard(s) for specific physical media
- Ð standard(s) for the migration of information across media and representations

1.5 DOCUMENT STRUCTURE

Section 2 provides a definition of the type of preservation to be achieved in an OAIS archive. It provides a high level view of what is meant by ÒinformationÓ in the context of the OAIS archive, and it gives a view of the environment of an OAIS archive. It also defines mandatory responsibilities an OAIS archive must discharge in preserving information.

Section 3 provides detailed model views of an OAIS archive. It breaks down the OAIS into a number of functional areas and it identifies some high level services at the interfaces It also provides detailed data model view of information using OMT diagrams.

Section 4 provides some perspectives on the issue or migration of information across media and across new formats or representations.

Section 5 provides some characteristics by which archives may be categorized.

Section 6 provides scenarios, using the terms and concepts defined previously, to show how information may flow into and out of an archive, and how information may be migrated within an archive.

Annex A provides scenarios of existing archive operations.

Annex B provides examples of information migration strategies..

Annex C relates parts of this reference model to other standards work.

Annex D provides a brief tutorial on the Object Modeling Technique (OMT)

1.6 DEFINITIONS

1.6.1 ACRONYMS AND ABBREVIATIONS

AIC - Archival Information Collection
AIP - Archival Information Package
ASCII - American Standard Code for Information Interchange
CCSDS - Consultative Committee for Space Data Systems
CD-ROM - Compact Disk - Read Only Memory
CIP - Catalog Interoperability Protocol
DBMS - Data Base Management System
DCP - Digital Collection Profile
DDL - Data Description Language
DIP - Dissemination Information Package
DR - Descriptive Record
EBCDIC - Extended Binary Coded Decimal Interchange Code (ck)
HFMS - Hierarchial File Management System
IEEE - Institute of Electrical and Electronic Engineers (ck)
ISO - Organization for International Standardization
LSDA - Life Sciences Data Archive
NARA - National Archives and Records Administration
NASA - National Aeronautics and Space Administration
NSSC - ???(find usage)
OAIS - Open Archive Information System
OMT - Object Modeling Technique
PDS - Planetary Data System
PI - Principal Investigator
SIP - Submission Information Package
UNICODE - Universal Code

WWW - World Wide Web

1.6.2 TERMS

Access - the services and functions which make the archival information and externally-available services visible to consumers, accept orders from consumers and provide customer services.

Active Archive- an archive where data is flowing regularly into the archive over an extended period of time rather than a single submission package. This data can also be accessed regularly by users.

Adhoc Consumer: Does not know what specific holdings of the archive they interested in acquiring.

Administration - This entity contains the services and functions needed to control the operation of the other Archive entities on a moment by moment basis

Archive Descriptive Records - records kept by the archive that track the day to day functioning of the archive including records regarding transactions, user support, etc.

Archives - Repositories that intend to preserve information for access and use by one or more designated communities.

Archived Information: Information, represented by digital or non digital data, that is being preserved for public or private access over the long (indefinite) term. The information is deemed to be understandable to one or more segments of the public. For the information to be preserved, the underlying representations may be changed as needed to maximize information preservation.

Archive Information Collections (AIC) aggregations of AIPÕs using criteria determined by the archivists in consultation with the information producers. Examples of AICÕs may be all the data and documentation from a single experiment, or all the data and documentation from a single region in space.

Archival Information Package (AIP) a set of information that has all the qualities needed for permanent, or indefinite-long term, preservation of a designated information object. An example of an AIP may be a spreadsheet of numbers representing temperatures in a certain region with all the associated documentation describing how the temperature was measured, what instruments were used to make the measurements, who made the measurements, why they were made, etc..

Catalog Information - information that is used by finding aids to allow users to locate objects of interest and that it is deemed desireable to preserve along with the Content Information.

Collection : An aggregation of two or more other objects, where those objects can be any combination of AIPs and/or Collections. Each Collection is considered to be suitable for being adequately documented for preservation, distribution and independent usage.

Collection Descriptive Records relevant metadata about the collection and pointers to contained information objects or collections.

Common Services - supporting services such as interprocess communication, name services, temporary storage allocation, exception handling, security, and directory services necessary to support the archive.

Consumer: Consumers play the role of those who interact with archive services to find information of interest and to access that information in detail.

Content Information- the primary information being preserved. An example of content information may be a single spreadsheet of numbers representing, and understandable as,

temperature but without the documentation which would explain its history and origin and how it relates to observations.

Context Information - This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere. An example of context information is information describing why the data was collected, i.e., objects, mission statement, etc..

Data : The representation forms of information. An example of data is

Data Delivery Session - A Data Delivery session may be a delivered set of media or a single telecommunications session. The Data Delivery session format/contents is based on a data model negotiated between the archive and the producer in the Submission Agreement. This data model identifies the logical constructs used by the producer and how they are represented on each media delivery or in the telecommunication session.

Data Dissemination Session -

Data Dictionary - a formal repository of terms used to describe data.

Data Management - the services and functions for populating, maintaining, and querying a wide variety of metadata including product related metadata such as data catalogs, directories, inventories, and processing algorithms and other metadata including information on customer access and security, archive schedules and procedures, and processing history.

Data Management Data - Data created and stored in a database that refer to operation of an archive. Some examples of this data are accounting data for customer billing and authorization, policy data, subscription data for repeating requests, and statistical data for

generating reports to archive management

Data Model: The collection of representations and their relationships which apply to a particular logical or physical data representation. [needs work]

Data Object either a Physical Object or a Digital Object.

Datastore - the set of all the portions of databases that make up a collection

Descriptive Records - records that describe information objects or collections that contain all the relevant descriptions for that object

Digital Collection Profile (DCP) - combines both a logical view of digital collections and a physical view of data collections.

Digital Library Profile - a companion profile that extends the DCP to include features needed for access to digital libraries.

Digital Object - an object composed of a set of bit sequences

Directories - a guide to what is contained in the archive (definition requested by CNES)

Dissemination - provides Dissemination Information Package to the Consumer

Dissemination Information Package - one or more AIP's that are distributed to the consumer per their request.

Finding Aid - is a tool provided to the user that allows a user to search for and identify potential holdings of interest.

Fixity Information - documents the authentication mechanisms and provides authentication

keys to ensure that the Content Information object has not been altered in an undocumented manner. An example of fixity information is CRC code for a file.

Format - The shape, size, style, and general makeup of a particular record or information object. (needs more tech definition CNES)

Independently Usable Information: Information with sufficient documentation to allow the data to be used by the designated user community without having to resort to special resources not widely available, including named individuals.

Information: Any type of knowledge that can be exchanged.

Ingest - a process whereby Submission Information Packages are accepted from the producers, and the packages are prepared for storage and management within the archive

Inventories - a record of what is contained within the archive or within certain domains of the archive. (requested by CNES)

Logical Data: Information representation forms that are NOT directly physically observable, but are inferred from established rules.

Long-term Preservation or Long-term Storage - long-term information preservation in which an "information package" containing both the information objects and the descriptions needed to interpret the information are potentially stored permanently .

Management - interacts with the Archive by providing policy guidelines and by receiving statistics relevant to evaluating adherence to the policy guidelines it has provided

Metadata : Data about other data.

Migration: moving from an older method of digital representation to a newer method.

Object Descriptive Records - records describing information objects that contain all the relevant metadata for that object.

Object Layer Information (Don will provide definition)

Permanent Data - the information that archive is tasked to preserve

Physical Data Object: Information representation forms that are physically observable properties.

Physical Objects - an object (such as a moon rock, biospecimen, microscope slide) that is considered suitable for being adequately documented for preservation, distribution and independent usage.

Preserve Information: Maintain information, in a correct and independently usable form over and indefinite period of time.

Preservation Descriptive Information - information necessary to adequately preserve the content information.

Producer - the entity that creates the information objects (either digital or physical) to be archived

Provenance Information - This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. An example of provenance information is the principal investigator who recorded the data and information concerning its storage, handling and migration.

Reference Information - This information identifies, and if necessary describes, one or

more mechanisms used to provide assigned identifiers for the Content Information. It also provides these identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. An example of reference information is a filename.

Reference Model: A reference model is a framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a non-specialist.

Representation Description - a description of a physical or digital object.

Representation Information - maps the physical bit level information into the content concepts addressed by the creator of the digital object, an example is the ASCII format which describes how the bits are represented.

Representation Methods - documentation that allows the interpretation of information objects from the bit level if the associated software becomes obsolete.

Representation Net - The set of Representation Information which fully describes the meaning of a Data Object. Representation Information in digital forms needs additional Representation Information so it can be understood.

Request Agreement an agreement between the archive and the customer in which the physical details of the delivery such as media type and object format are specified.

Results Set - the set of descriptive records for those AIPs in an OAIS which match the criteria stated in a user/consumer query.

Search Session - a session initiated by the consumer with the archive in during which the

consumer will use the archive finding aids to identify and investigate potential holdings of interest

Security - (Lou will define).

Semantic Layer Information - (see Object Layer Information)

Storage - provides the services and functions for the storage and retrieval of data objects, both logical and physical, including non-digital media.

Structure Layer Information - translates the bit streams to common computer types such as characters, numbers, and pixels and aggregations of those types such as character strings and arrays.

Submission Agreement - an agreement reached between an archive and the producer that negotiates a data model for the Data Delivery session. This data model identifies format/contents and the logical constructs used by the producer and how they are represented on each media delivery or in the telecommunication session. It also transfers legal and physical custody of the data to the archive and the terms defining restrictions and access.

Submission Information Package - the information package identified by the Producer in the Submission Agreement with the OAIS

Subscription Consumer - establishes a Request Agreement (i.e. a subscription) with the archive which may span any length of time with one or more Data Dissemination sessions usually with significant time gaps between the sessions agreed upon.

Traditional Archive - archives that deal with paper, photographic, or other records typically funded by government, organizations, institutions, or corporations.

Transformation - the movement of data from one format to another without losing

information or changing the content information which you are trying to preserve.

Transient Data - data which is produced, used, and discarded in day-to-day business operations

1.7 REFERENCES

[1] "Preserving Digital Information", Report of the Task Force on Archiving of Digital Information, final report currently available from the URL: http://www.rlg.org/ArchTF/ ,May 1, 1996

[2] "Object-Oriented Modeling and Design", by Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W., Prentice Hall, 1991.

[3] "Z39.50 Profile for Access to Digital Collections", http://vinca.cnidr.org/protocols/profiles/zdl.html

[4] "Z39.50 Profile for Access to Digital Library Objects", http://lcweb.loc.gov/z3950/agency/profiles/dl.html , August 15, 1996.

2 OAIS CONCEPTS

2.1 OAIS FUNCTION

The term ÒarchiveÓ has come to be used to refer to a wide variety of storage and preservation functions and systems. Traditionally, an archive is understood as a facility or organization which preserves records, originally generated by or for a government organization, institution, or corporation, for access by public. or private communities. It accomplishes this task by taking ownership of the records, ensuring that they are understandable to the accessing community, and managing them so as to preserve their information content and authenticity. Historically, such records have been in such forms as books, papers, maps, photographs, and film which can be read directly by humans, or read with the aid of simple optical magnification and scanning aids. The major focus for preserving this information has been to ensure that they are on media with long term stability and that access to this media is carefully controlled.

The explosive growth of information in digital forms has posed a severe challenge not only for traditional archives and their information providers, but for many other organizations in the commercial and private sectors. These organizations are finding, or will find, that they need to take on the information preservation functions typically associated with traditional archives because digital information is easily lost or corrupted. The pace of technology evolution is causing some hardware and software systems to become obsolete in a matter of a few years, and these changes can put severe pressure on the ability of the related data structures or formats to continue effective representation of the full information desired.

A major purpose of this reference model is to facilitate a much wider understanding of what is required to preserve information permanently or for the indefinite long term. To avoid confusion with a "bit storage archive", the reference model defines an Open Archival Information System (OAIS) which performs an archival function. An OAIS archive is one that meets the minimum requirements given in section 2.4, and when future OAIS standards are defined, conforms to them as well. This Archival Information System is considered to be "Open" because the model and its standards are being developed using a public process and are readily available to the public. "Open" does NOT mean that access to information within the archive is uncontrolled. For the remainder of this document, the term archive is understood to refer to an OAIS, or OAIS archive, unless the context makes it clear otherwise (e.g., traditional archives).

The OAIS model recognizes the already highly distributed nature of digital information holdings and the need for local implementations of effective policies and procedures supporting information preservation. This allows, in principle, a wide variety of organizational arrangements, including various roles for traditional archives, in achieving this preservation. It is expected that organizations, both large and small, attempting to preserve information, particularly digital information, will find that conforming to OAIS requirements and standards, and using OAIS terms and concepts, will help them achieve their information preservation goals.

At the same time, the problems associated with achieving economical and truly effective permanent or indefinite long term digital information preservation should not be underestimated. A good survey of many of the issues is contained in a report by the Task Force on Archiving of Digital Information entitled ÒPreserving Digital InformationÓ [Reference 1].

2.2 OAIS INFORMATION

A clear definition of information is central to the ability of an OAIS to preserve it. The intent of this section is to begin to give a framework for recognizing the various types of information involved. However the mapping of specific cases into such models often depends on local archive considertions and this makes the topic more complex. Example mappings are given to help understand the concepts involved. Further details on information modeling are given in Section 3.2.

An **Archival Information Package** (AIP) is defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite-long term, preservation of a designated **information object**. The AIP is itself an information object that is a container of other information objects. Within the AIP is the

designated information object and it is called the **Content Information**. Also within the AIP is an information object called the **Preservation Description Information**. The Preservation Description Information provides additional information about the Content Information and is needed to make the Content Information meaningful for the indefinite long-term. These relationships are modeled in Figure 2-1 using the **Object Modeling Technique** (OMT) [Annex D and Reference 2]. As might be expected, an OAIS must clearly understand what constitutes the Content Information for each specific case in order to ensure that the corresponding Preservation Description Information fully performs its function.

Figure 2-1: Archival Information Package (AIP)

2.2.1 CONTENT INFORMATION

The Content Information is that information which is the primary object of preservation. The OAIS needs to explicitly decide what this information is in order to be able to ensure that it also has the necessary Preservation Description Information which is needed to preserve the Content Information. Deciding what is the Content Information, for a given set of information, may not be obvious and may need to be negotiated with the information producer.

For example, consider an OAIS that archives documents of other government organizations. A 10-page, hard copy, document was generated by some Government organization to describe its services and products and is submitted for preservation. Assume that it is determined that the primary information to be preserved is the description of the services and products, including its organization on each page. It is determined that it is NOT important to preserve the actual pages themselves. The Content Information, in this case, would be the descriptions of the services and products and the way this information is organized on each page. This Content Information could, in principle, be moved to new media, including electronic media, without serious risk of Content Information loss because the original hardcopy pages have been determined to be outside this preservation effort. In general, there are layers, or levels, of information in every information object and the OAIS needs to understand which layers constitute the Content Information.

As another and more complex example, consider an electronic file containing a sequence of values obtained from a sensor looking at the Earth's environment. There is a second file, encoded using ASCII, that provides information on how to understand the first file. It describes how to interpret the bits of the first file to obtain meaningful numbers, it describes what these numbers mean in terms of the physics of the observation being conducted, it gives the date and time period over which the observations were made, it gives an average value for the observed values, and it describes who made the observations. These two files are submitted to an OAIS for preservation.

Assume that the OAIS determines that the Content Information to be preserved is the observed bits together with their values as numbers and the physics meaning of these numbers. This information is conveyed by the bit sequence within the first file together with the **Representation** information from the second file needed to transform the first fileÕs bits into meaningful physical values. Note that neither the first fileÕs underlying media nor the particular file system carrying the bits is part of the Content Information. From the second file only part of its content is considered a part of the Content Information and this is the part that enables the transformation of the bits from the first file into meaningful physical values. In fact this second file does not carry

all the Representation information needed to make this transformation because the following additional information is needed:

- Information that the second file is encoded in ASCII so that it can be read as meaningful characters;

- Information on how the characters are used to express the transformations from bits to numbers to meaningful physics values. This information, typically referred to as a combination of format information and data dictionary information, may also include instrument calibration values and information on how the calibrations are to be applied. All this information may be widely understandable once the ASCII characters are visible because it has all been expressed in English (or some other natural language), or some of it may be in more structured forms that will need additional Representation information to be understood.

Therefore the Representation information of the second file needs additional Representation information, and this information may need additional Representation information, etc., forming a linked set of Representations of Representations. Because each Representation can be composed of multiple components, each with is own Representation, the result can be described as a **Representation Net**. All this additional Representation information is needed to fully transform the bits of the first file into the Content Information. In principal, this even extends to the inclusion of definitions (e.g., dictionary and grammar) of the natural language used (English in this example). Over long time periods the meaning of natural language expressions can evolve significantly in both general and in specific discipline usage. As an aside, it is clear from this example that, in general, 100% information preservation for the indefinite long-term is a goal that is not practically achievable.

As a practical matter, the OAIS needs to have enough Representation information associated with the bits of the first file that it feels confident that it, and those expected to use the information, can enter the Representation Net with enough knowledge to begin accurately interpreting the Representation information. This is a significant risk area for

an OAIS, particularly for those with a narrow discipline focus, because jargon and apparently widely understood terms may subsequently be found to be quite temporary.

In the example under discussion, the Content Information consists of some of the information in the two electronic files together with some information outside the files. This can be viewed as consisting of a primary **Digital Object** (the identified sequences of bits within the first file) and this digital objectÕs Representation Information (some information (bits) fom the second file and some information from outside this file). More generally, the Content Information can be viewed as a primary **Data Object** together with its Representation information as shown in Figure 2-2. The primary Data Object may be either a Digital Object or a Physical Object (e.g., a physical sample, microfilm). In this latter case the Representation information provides interpretations of physical attribues of the Physical Object to convey the intended Content Information

Figure 2-2: Content Information

Recall that in the second example above, there was a determination that the Content Information consisted of the observed sensor values and their meanings. This is by no means the only determination that could have

been made. It could just as easily have been determined that the Digital Object of the desired Content Information was the the bit sequences within the first file together with the all the bit sequences within the second file. The fact that some of these latter bit sequences are used to interpret the first files bit sequences is irrelevant and is just an example of a set of bits that is somewhat self-describing. It is also irrelevant that some of the bits in the second file are the basis for information on the date and time period over which the observations were made, the average value for the observed values, and who made the observations. Once it has been determined that all these bits constitute the Digital Object of the Content Information, then the Representation Information is that information needed to turn them into meaningful information. How extensive this meaning is to be carried and how far the Representation Net needs to be carried are local issues for the OAIS and its related producer and consumer communities. This OAIS environment is discussed further in Sections 2.3 and 3.

2.2.2 PRESERVATION DESCRIPTION INFORMATION

Once the Content Information has been determined, it is possible to assess thePreservation Description Information. The Preservation Description Information is that information which is necessary to adequately preserve the particular Content Information with which it is associated. It is specifically focused on describing the past and present states of the Content Information, ensuring it is uniquely identifiable, and enssuring it has not been unknowningly altered. This information can be categorized as follows:

- **Provenance Information**: This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This give future users some assurance as to the likely reliability of the Content Information. This information may be thought of as a special case of Context Information described below.

- **Reference Information**: This information identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Information.

Ð **Context Information**: This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere.

Ð **Fixity Information**: This information documents the authentication mechanisms, and it provides any authentication keys used to ensure that the particular Content Information object has not been altered in an undocumented manner.

Ð **Catalog Information**: This optional information has been extracted from the Content Information to assist in finding the Content Information when searches are made. To avoid the possibility of having to re-extract the information in the future, it is added to the Preservation Description Information set. This information may be thought of as a particular type of Reference Information.

Figure 2-3: Preservation Description Information

2.2.3 INFORMATION PACKAGE VARIANTS

It is necessary to distinguish between an AIP that is preserved by an OAIS and the information packages that are submitted to, and disseminated from, an OAIS. These variant packages are needed to reflect the reality that some submissions to an OAIS will have insufficient Representation information or Preservation Description Information to meet final OAIS preservation requirements. In addition, they may be organized very differently from the way the OAIS organizes the information it is preserving. Finally, the OAIS may provide information to consumers that does not include all the Representation information or all the Preserving Description Information with the associated primary Data Objects being disseminated. These variants and the types of transformations among them need to be described.

The variants of the AIP (in addition to the AIP itself) are modeled as subclasses of a new class called **Information Packages** as shown in figure 2-4.

Figure 2-4: Information Package Object Diagram

The Submission Information Package (SIP) is that package that is sent to an OAIS by a data producer. Within the OAIS it is transformed into an **Archival Information Package** (AIP) for preservation. In response to a request, the OAIS provides all or a part of an AIP

in the form of a **Dissemination Information Package** (DIP). The OAIS environment is described in the following section.

2.3 OAIS ENVIRONMENT

The environment surrounding an OAIS is given by the simple model shown in Figure 2-5

Figure 2-5: Environment Model of an OAIS

Outside the OAIS are Producers, **Consumers**, and **Management**. Producers play the role of those who provide information to be preserved. Management sets OAIS management requirements that are consistent with a management environment need in which the OAIS is only one of its responsibilities. Management is not involved in day-to-day archive operations. This functionality is included within the OAIS. Consumers play the role of those who interact with OAIS services to find information of interest and to access that information in detail.

Other OAIS archives are not shown explicitly however such archives may establish particular agreements among themselves consistent with Management and OAIS needs. Other archives may interact with a particular archive for a variety of reasons and with varying degrees of formalism for any pre-arranged agreements. They may act as producers

such as when responsibility for preserving a type of information is to be moved to another archive. They may act as consumers and rely on a different archive for a type of information they seldom need and chose not to preserve locally. Such arrangements should have some formal basis, requiring communication, to help ensure that access to the needed information is not lost when the policies of an OAIS change.

The following sections present a high level view of the interaction between the high level entities in the OAIS environment. Figure 2-6 is a data flow diagram that represents the operational OAIS archive external data flows. This diagram concentrates on the flow of information among producers, consumers and archives in the OAIS environment and does not include flows that involve high level management.

Figure 2-6 OAIS Archive External Data Flows

2.3.1 PRODUCER INTERACTION

The Producer establishes a **Submission Agreement** with the OAIS, which identifies the Submission Information Packages (SIPs) to be submitted and may span any length of time

for this submission. Within the Submission Agreement, one or more **Data Delivery Sessions**, usually with significant time gaps between the sessions, are recognized. A Data Delivery session may be a delivered set of media or a single telecommunications session. The Data Delivery session content is based on a data model negotiated between the OAIS and the Producer in the Submission Agreement. This data model identifies the logical components of the SIP that are to be provided and how they are represented on each media delivery or in the telecommunication session. All Data Deliveries within a Submission Agreement are recognized as belonging to that Agreement and will generally have a consistent data model which is specified in the Submission Agreement. For example, a data delivery session might consist of a set of Content Information objects corresponding to a set of observations which are carried by a set of files on a CD-ROM. The names of the files or the directory structure of the CD-ROM might be used to store information about the observation times or the datatypes contained in each file. The Submission Agreement would detail the file format and the convention for the filenames along with the frequency of data delivery sessions (e.g. one per month for two years). It would also give other needed information such as access restrictions to the data and it would state how the associated Provenance, Context, and Reference information were to be provided.

Each SIP in a Data Delivery session is expected to meet minimum OAIS requirements for completeness so that it may be fully ingested into the OAIS archive. This means that the information contained therein can become available, in principle, to OAIS Consumers. A Submission Agreement also includes the procedure and protocols by which an OAIS will either verify the arrival and completeness of a data delivery session with the producer or question the producer on the contents of the data delivery session.

2.3.2 CONSUMER INTERACTION

There are two basic classes of Consumers, the Subscription Consumer and **the Adhoc Consumer**. The Subscription Consumer establishes a **Request Agreement** (i.e. a subscription) with

the OAIS which may span any length of time. Within the Request Agreement, one or more **Data Dissemination Sessions**, usually with significant time gaps between the sessions, take place. A Data Dissemination session may involve the transfer of a set of media or a single telecommunications session. The Data Dissemination Session contents is based on a data model negotiated between the OAIS and the consumer in the Request Agreement. This data model identifies the components of one or more AIPs to be provided, how they are mapped into a Dissemination Information Package (DIP) and how that DIP will be represented on each media delivery or telecommunications session. The Request Agreement will also specify other needed information such as the trigger (e.g. event or time period) for new Data Dissemination sessions, the criteria for selecting the OAIS holdings to be included in each new Data Dissemination session, delivery information (e.g., name or mailing address), and any pricing agreements.

The Adhoc Consumer does not know a priori what specific holdings of the OAIS he is interested in acquiring. The Adhoc consumer will establish a **Search Session** with the archive. During this Search Session an Adhoc Consumer will use the archive finding aids to identify and investigate potential holdings of interest. This searching process tends to be iterative with a user first identifying broad criteria and the refining his criteria based on previous search results. Once the Adhoc Consumer identifies the archive AIP components he wishes to acquire, he must establish a Request Agreement with the archive to establish the details of the how he will acquire these desired components. At this point the Adhoc Consumer will act as a Subscription Consumer though the Request agreements may be substantially simpler than those negotiated with Subscription Customers. It should be clear that any consumer can act as a Subscription Consumer and/or an Adhoc consumer based on his current needs and requirements. Note that the concept of a Request Agreement does not specify any particular implementation. It may, in some cases, be no more than the completion of a WWW form as to what is desired.

2.3.3 MANAGEMENT INTERACTION

Management provides the OAIS with its charter and scope. The charter may be developed by the archive but it

is important that management formally endorse archive activities. The scope determines the breadth of both the producer and consumer groups served by the archive. Management is often the primary source of funding for an archive and may provide guidelines for resource utilization (personnel, equipment, facilities). Management will generally conduct some regular review process to evaluate archive performance and progress toward long-term goals. Management determines or at least endorses pricing policies for archive services. Management participates in conflict resolution involving producers, consumers and archive administration. Management should also provide support for the archive by establishing procedures that assure archive utilization within its sphere of influence.

2.4 OAIS RESPONSIBILITIES

This section establishes mandatory responsibilities that an organization must discharge in order to operate an OAIS archive.

- Negotiates and accepts Submission Information Packages from information producers

- Determines (dependently or independently) which communities need to be able to understand the Content Information objects provided.

- Ensures the information, to be preserved, is independently understandable to the designated communities. In other words, the communities should be able to understand the information without needing the assistance of the experts who produced the information.

- Assumes sufficient control of the information provided to the level needed to ensure permanent or indefinite long-term preservation.

- Follows established policies and procedures which ensure the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as traceable to the original.

Ð Makes the preserved information available to the designated communities in forms understandable to those communities.

In the following sections, each of these responsibilities is explored in greater depth for clarification.

2.4.1 NEGOTIATES AND ACCEPTS SIPS

An organization operating an OAIS will have established some criteria that aids in determining the types of information that it is willing to, or it is required to, accept. These criteria may include, among others, subject matter, information source, degree of uniqueness or originality, and the nature of the techniques used to represent the information (e.g., physical media, digital media, format). The information may, in general, be submitted using a wide variety of common and not-so-common forms, such as books, documents, maps, data sets, and moon rocks using a variety of communication paths including networks, mail, and special delivery.

To further extract some commonalty across these forms, the reference model has defined the concept of an "Archival Information Package." The AIP provides a conceptual basis around which to formulate more specific archival policies on what is acceptable in a SIP. It requires a clear distinction be made about what constitutes the information that is to be preserved (i.e., Content Information). As described in Section 2.2, for the exact same Content Information object, different decisions can be made about which of the provided levels of information should be preserved. The AIP also requires a clear distinction about what constitutes the necessary associated description information (i.e., Preservation Description Information). These categories of information ultimately need to be present even if they are not provided in a SIP with each Data Delivery Session.

Ideally, the archive and the data producer, who is submitting information to the archive, will agree in advance on what the SIPs should be and how they will be delivered as documented in the Submission Agreement. This will facilitate the ingest of this information into the archive. However, the archive may need to redefine what are appropriate AIPs, corresponding to a set of provided SIP information, in order to provide effective customer services. This may also include the need to define various Collections of AIPs (i.e., CIPs). The modeling of such collections is discussed in Section 3.2.

For example, a 10 page manuscript may be made available as an SIP. Subsequently, each page may be made independently findable using an archive generated index. If each page

is also documented with appropriate Preservation Description Information, then each page becomes an AIP and the manuscript can be documented as a Collection of AIPs.

As another example, consider a month's worth of scientific observations of the magnetic field components in the interplanetary medium. Observations are made every 10 seconds and stored in a record. The records are accumulated for a day and put into a file. Thirty files comprise the month's set of observations. Initially, the information may be submitted with the view that each file, along with its Representation Information, will be a Content Information object. If each file is adequately documented with Preservation Description Information, then this will constitute an AIP and each day's worth of observation will be individually findable and readily presentable to a customer by the archive. It will be up to each customer to do further searching within the file content, or Content Information, for more specific information. The month's worth of observations can be documented as a Collection. Later, the archive may decide that it should provide the ability to find and extract individual 10 second observations. It builds the necessary index or algorithms to do this search. The individual records, together with their Representation Information, become Content Information presentable to customers in the context of the daily AIPs. If these information objects are subsequently documented with Preservation Description Information, they form new Ò10-second ÒAIPs. Each file can be documented as a collection of 10-second AIPs, and the month's worth of observations can be a collection of the collection of 10-second AIPs.

2.4.2 DETERMINES DESIGNATED CONSUMER COMMUNITIES

The submission, or planned submission, of an SIP requires a determination as to who the expected consumers of this information will be. This is necessary in order to determine if the information, as represented, will be understandable to that community.

For example, an archive may decide an SIP Content Information object should be understandable to the general public. It will need to be sure that all documentation, expected to be read by the general public, is free of

jargon and is widely understandable.

For some scientific information, the designated community of consumers might be described as those with a first year graduate level education in a related scientific discipline. This is a more difficult case as it is less clear what degree of specialized scientific terminology might actually be acceptable. The producers of such specialized information are often used to a narrowly recognized set of jargon, so it is especially critical to clearly define the designated community for their information and to make the effort to ensure this community can understand the information. If the archive does not have this level of expertise in-house, it may need to have outside community representatives review the information for long-term understandability.

2.4.3 ENSURES INFORMATION IS INDEPENDENTLY USABLE

The degree to which an AIP conveys information to a designated community is, in general, quite subjective. Nevertheless, it is essential that an archive make this determination in order to maximize information preservation.

For example, a manuscript Content Information object may be written in English and therefore its content may be generally understandable to a wide audience. However, unless the purpose for which it was created is clearly documented, much of its meaning may be lost. This 'purpose' information is part of its current context that must be provided in the Preservation Description Information.

As another example, consider again SIPs from the set of scientific observations of the interplanetary magnetic field. Typically such information is not in a form intended for direct human browsing or reading, but is rather in a form appropriate to searching and manipulation by application software. Such content may only be understandable to the original producers unless there is adequate documentation of the meaning of the various fields and their inter-relationships, and how the values relate back to the original instrumentation that made the observations. In such specialized fields extra effort is needed to ensure that the Content Information and the Context Information are understandable to a designated community. Otherwise the information may be understandable to only a few specialists and be lost when they are no longer available.

Digital Content Information objects need software for efficient access. However, maintaining Content Information object-specific software over the long term has not yet been proven cost effective.

In general, each Content Information object should have its information representations fully documented down to the bit level, even if the lower levels may, at ingest time, be supported by widely available software. Over time, all representations will be replaced and documenting them ensures that they are recognized by the archive

and can be tracked as to viability for usage by the archive customers. This also facilitates Content Information object migration to new representations.

2.4.4 ASSUMES SUFFICIENT CONTROL FOR PRESERVATION

Upon acceptance of a SIP/AIP, or collection of AIPs, the archive must assume sufficient control over the information so that it is able to preserve it over the long term.

For example, consider a digital Content Information object where the highest levels of meaning in the Representation Information are those to be preserved. Unless the archive also has control of the underlying data structures and is able to evolve them to keep pace with widely accepted standards, effective access to the information could be greatly restricted over time. The archive must be fully cognizant of the information that it is to preserve, and be able to change its representation as is most appropriate. At the same time, this does not preclude retaining all original data structures as historical information that is traceable through the Provenance Information. This Provenance Information would be updated to reflect such migrations within the archive.

Archives that need to preserve copyrighted digital material will need to understand exactly what levels of information the copyright applies to so that they can retain maximum flexibility in performing indefinite long-term preservation.

2.4.5 FOLLOWS ESTABLISHED PRESERVATION POLICIES AND PROCEDURES

It is essential for an archive to have established policies and procedures for preserving its AIPÕs, and to follow those procedures. This is particularly true for digital AIPÕs or digital Content Information objects because of their frequent need for attention as discussed earlier. They are also highly vulnerable to inadvertent and intentional destruction or corruption even in the most trusted systems. This fragile nature puts a premium on being able to recover from all digital AIP handling operations when errors are discovered.

The appropriate policies and procedures will depend, at minimum, on the nature of the AIPÕs and any ÔbackupÕ relationships the archive may have with other archives.

In general, digital SIPÕs received may undergo some transformations during the ingest process before being fully incorporated into the OAIS as AIPs, may undergo migrations with transformations/revisions while in the OAIS, and may undergo transformations upon dissemination as DIPs to customers. For example, one extreme on ingest is to extract all relevant ÔvaluesÕ from the Content Information objects (thus forming new Content Information objects) as submitted and store them according to a complex schema, or data model, covering the archiveÕs subjects of interest. Such a transformation should be fully documented and traceable to the original SIPÕs. During internal migrations, some or all of the representations used to carry the information of an AIP may be replaced. These new representation need to be carefully documented and the transformation process needs to be fully described. Upon dissemination of an AIP, or part of an AIP, to a customer, a new representation may be used to provide more effective usage. Again, this transformation needs to be fully described and the descriptions of all past transformations need to be available to the customer. This attention to detail, while also ensuring against processing errors, requires that strong policies and procedures be in place and that they be executed.

A long term technology usage plan, updated as technology evolves, is also essential to avoid being caught with very costly system maintenance, emergency system replacements, and costly data representation transformations.

2.4.6 MAKES THE INFORMATION AVAILABLE

The expectations of OAIS customers will vary widely among archives and over time as technology evolves. By definition, an OAIS makes its AIPÕs and Collections visible and available to customers. Multiple views, supported by various search aids that cut across Collections, may be provided. Some AIPÕs, visible to customers, may not be stored as AIPÕs in any recognizable sense. Rather, they may be generated upon request using an associated algorithm operating on one or more existing AIPÕs. Upon dissemination, these derived AIPÕs will need to include documentation on how they were derived from other AIPÕs. For security, direct access by customers to stored AIPs should only be allowed on

copies of the original AIPs.

In general, AIPÕs and Collections will be distributed by all varieties of communication paths, including networks and physical media.

3 Detailed Models

THE PURPOSE OF THIS SECTION IS TO PROVIDE MORE DETAILED MODEL VIEW OF THE FUNCTIONAL AREAS OF THE OAIS AND THE INFORMATION HANDLED BY THE OAIS.

3.1 FUNCTIONAL MODEL

The OAIS of Figure 2-1 is broken into seven functional areas and related interfaces as shown in Figure 3-1. The lines connecting functions identify communication paths over which information flows in both directions. No connections are shown for Common Services to avoid clutter as this function communicates with all other functions. Similarly, no internal connections are shown to Administration because this function also communicates with all other functions.

Figure 3-1: OAIS Functional Entities

The role provided by each of these entities is briefly described as follows:

3.1.1 COMMON SERVICES

Modern, distributed computing applications assume a number of supporting services such as interprocess communication, name services, temporary storage allocation, exception handling, security, and directory services. This entity provides a single repository for these services.

3.1.2 INGEST

This ingest entity provides the services and functions to accept and validate a submission information package and prepare the contents for storage and management within the archive. In summary there is a *scheduling* function to negotiate delivery of the submission information package, the package is physically received by the archive (*staging*); the submission information package is *reviewed* by the archive staff and others; liens are *reported* to the data preparer; data *conversion* functions are applied as necessary to comply with internal archive representation formats; metadata and special products are *extracted* from the package for incorporation in the data management system; metadata is *created* to document and enhance the utility of the package; the archival information package is transferred (*transfer initiation*) to storage. Status reporting to both the data preparer and the archive administration are executed by the *ingest reporting* function. The ingest functions are described in more detail below.

* Scheduling. The scheduling function will negotiate data submission schedules with the data preparers and maintain a calendar of expected submission information packages and resource requirements to support their ingest.

* Staging. The staging function makes physical storage space available for a data submission session. The archive staff must allocate the appropriate storage capacity or devices to the data preparer. The data is delivered via electronic transfer (e.g. ftp) to the staging area; loaded from media submitted to the archive; or simply mounted (e.g. CD-ROM) on the archive file system for access. The staging function may represent a legal

transfer of custody for the products in the submission information package and may require that special access controls be placed on the contents of the submission information package.

* Review. The review function provides a validation of the submission information package and assures that the information is understandable to the intended consumer community. The review may be carried out by the archive data engineers and may also involve an outside committee (peer review). The review must verify that: 1) the data has been physically transported correctly to the archive staging area; 2) the quality of the data meets the requirements of the archive or peer review group; 3) the documentation and metadata are sufficient to make the data understandable to the target user community. The formality of the review will vary depending on internal archive policies. The review process may determine that some portions of the submission information package are not appropriate for inclusion in the archive and must be redone or excluded. This notification is carried out by the ingest reporting function.

* Conversion. The conversion function transforms the submitted information package into an archive information package that conforms to the internal data standards of the archive. It should be a high priority of the archive to see that the majority of submission information packages follow archive format and content standards as specified in a data submission manual. Coversions that modify the representation of data items (e.g. changing floating point representations) must be carefully montored and tested to assure the integrity of the converted data. The archive is assumed to have a pool of standard utilities to support the conversion of submitted data or metadata objects to internal format.

* Extract metadata. The extract metadata function gathers metadata values from the submitted objects for loading in the archive catalog. These metadata values are used to generate higher-level summaries and finding aids.

* Create metadata. The create metadata function generates metadata needed to support one

or more access views to the archive data management data base. This may include the production of special versions of a data object to support on-line access (thumbnail or browse images, for example), or production of special formats of data products or documentation (e.g. Acrobat PDF files).

* Transfer initiation. The transfer function moves the archival information package from the staging area to the storage area and provides metadata to the data management function for cataloging. This may be either an electronic or physical transfer. The transfer of metadata to the data management entity should include the set of transactions required to update the data management data base . The transfer of data objects to the storage entity will include a storage request detailing the location, media type and volume of data to be stored.

* Ingest Reporting. The ingest reporting function provides interaction between the archive staff and the data preparer. The initial confirmation is an acknowledgment of receipt of a submission package. After the review process is completed any liens are reported to the preparer who will then resubmit or appeal the decision. After review completion a final report on the submission session is prepared for distribution to the administration and to the producer.

3.1.3 ARCHIVAL STORAGE

This entity provides the services and functions for the storage and retrieval of archive information packages and data objects that comprise them. Archival storage functions include *transferring* data from staging storage to permanent storage; managing the *storage hierarchy*; migrating data to new media over time; performing routine and special *error checking*; providing automated *backup* procedures; producing *duplicate* copies of portions of the archive; and *reporting* on storage activities.

* Transfer receipt. The transfer function receives a transfer request of an archive information package from staging storage and moves the data to permanent storage within

the archive. The transfer request should indicate the anticipated frequency of utilization of the data objects comprising the archive information package to allow the appropriate storage devices or media to be selected for storing the archive information package. The transfer function will select the media type, prepare the devices or volumes and perform the physical transfer to the storage volumes.

* Hierarchy management. The hierarchy management function positions data objects and collections on the appropriate media based on usage statistics or access requirements determined by administration. Hierarchy management procedures must follow policies for the use of on-line versus near-line versus off-line storage. To the extent possible the hierarchy management function should be automated based on usage statistics generated by the storage reporting function.

* Physical Migration. The physical migration function is responsible for maintaining the validity of the data objects across media over time. This migration does not change the underlying representation information. The migration strategy must take into consideration the expected and actual rates of error encountered in various media types and assure that data objects are converted with minimal data loss. The migration function must find a way to maintain the unique attributes of various media types (e.g. tape block sizes, CD-ROM volume information) when migrating to higher capacity media with different storage architectures.

* Error checking. The error checking function provides statistically acceptable assurance that no data objects are corrupted during transfer, migration or backup procedures. This function requires that all hardware and software within the archive provide notification of potential errors and that these errors are routed to standard logs that are checked by the storage staff. A standard mechanism for tracking and verifying the validity of all data objects within the archive should also be used. For example, cyclical redundancy checks (CRCs), should be maintained for every individual data file. The storage facility procedures should provide for random verification of the integrity of data objects using

CRCs or some other error checking mechanism.

* Backup. The backup function provides an automated mechanism for producing a duplicate copy of the archive contents. The backup media should be capable of being physically removed from the archive for storage at a separate facility. Backup requirements are specified in the archive procedures for various types of data.

* Duplicate. The duplicate function provides copies of stored data objects for distribution on physical media to requesters on media types supported by the archive. The duplicate request must provide an itemized list of data objects or physical volumes and identify the output media type. The duplicate function must prepare the output media, perform the transfer from storage to the new media and provide logical and physical labels for distribution with the duplicate copy.

* Storage Reporting. The storage reporting function provides regular reports to administration summarizing the inventory of media on-hand, available storage capacity in the various tiers of the storage hierarchy, and operational statistics.

3.1.4 DATA MANAGEMENT

This entity provides the services and functions for populating, maintaining, and querying a wide variety of metadata including product related metadata such as data catalogs, directories, inventories, and processing algorithms. It also provides access to other metadata including information on customer access and security, archive schedules and procedures, and processing history. The data management functions include controlling *access* to the data management system; providing services for *requesting* and *generating* reports; providing the capability to *update* the data base; allowing custom *schema definitions* and *view definitions* for users and potentially using *data mining* techniques to extract further information from the data managment data base. This entity includes all information needed (excepting archival storage) for archive operations.

* Report Request. The request function provides the capability to specify the contents and logical criteria for generating customized report for internal or external use. It must

provide the capability to store report requests and to generate periodic reports or reports triggered by logical criteria on a periodic basis.

* Report Generation. The report generation function will execute a report request and transmit the output to the user.

* Update. The update function accumulates updates to the data management data base and applies these transactions on a periodic basis, producing a summary of all modifications made to the data management data base (audit trail).

* Metadata maintenance. The metadata maintenance function is responsible for reviewing and updating metadata values (e.g. contact names, and addresses) on a periodic basis.

* Data base Administration. The data base administration function is responsible for maintaining the integrity of the data base; for creating any schema or table definitions required to support data management functions; and for providing the capability to create, maintain and access customized user views of the contents of the database.

3.1.5 ADMINISTRATION

The administration function shall manage all of the system activities. These include data *acquisition* efforts, maintaining *configuration management* of system hardware and software, *planning and scheduling* archive facility resources, performing *accounting* functions to bill users for services; providing *customer service* functions to users and performing *data engineering* work to develop and maintain archive standards. .

Acquisition. The acquisition function shall solicit additional data for inclusion into the system and shall handle administrative aspects of acquiring new data sets.

Configuration Management. The configuration management function shall maintain configuration control over the archive system, systematically controlling changes to the configuration. This function shall also maintain integrity and traceability of the configuration during all phases of the system life cycle.

Physical Access Control. The physical access control function will provide mechanisms to restrict or allow physical access (doors, locks, guards) to elements of the archive as determined by archive policies.

Planning and scheduling. The planning and scheduling function shall schedule system usage, times of heavy resource utilization, system down times for maintenance, and system upgrades.

Monitoring. The monitoring function shall audit system operations, system performance, and system usage.

Accounting. The accounting function shall bill and collect users for the utilization of archive system resources.

Customer Service. The customer service function shall provide any necessary assistance to archive system users. This service shall include answering questions, resolving user problems, providing information about and documentation on the system, providing status of orders, and about the status of data ingest activities. This function shall also create, maintain and delete user accounts.

Data Engineering. The data engineering function supports all ingest functions and is responsible for developing and maintaining the archive system data standards. The data submission formats and procedures must be clearly documented in the archiveÕs data submission manual and the deliverables identified by the preparer in the submission agreement..

3.1.6 ACCESS

The access function shall support the user in determining the existence, description, location and availability of information of interest. The function shall allow the user to view an *overview* of the contents of the archive or to *query* the database directly; *retrieve* selected data objects; to *manipulate* those objects using system software resources, to *display* the processed data interactively; to *order* selected data sets for on-line or off-line delivery; and to continue to *develop* new access capabilities. These functions shall be applicable across the range of catalog data, sample and summary data, and archival data which is in machine-readable format within the archive system.

* Access Control. The access control function provides system security. A hierarchy of security controls may be implemented depending on the needs of the archive system. These include limiting physical access to the system, establishing firewalls to prevent communication outside an area, electronic signatures and authorization procedures, restricting access to certain network domains, and assignment of user names and passwords. Any combination of these procedures may be needed in certain archive scenarios.

* Overview. The overview (browse) function provides an overview of data sets available in the archive system. In general the browse capability will provide summary versions of products which can be quickly viewed (thumbnails of images, abstracts of documents, etc.).

* Query. The query function provides the capability for the user to retrieve information from the data base management system by specifying Boolean search criteria.

* Retrieve. The retrieve function provides access to data which is on line, or requests the staging of data which is not available for immediate inspection because it is currently off line.

* Manipulate. The manipulate function provides a suite of functions which allow the user to perform processing operations on retrieved data (e.g. statistical,

* Display. The display function provides a predefined suite of functions for displaying retrieved data in textual, graphic, or image form. In addition, all data responses from the inspect data function to the user shall pass through this function.

* Order. The order function shall provide all services required to accept an order from a user, insure its validity, confirm the reasonability of the order from the viewpoint of the user, verify that all required information has been provided, and prepare the order for its actual execution by the dissemination function.

* Advanced Development. The advanced development function is responsible for continually updating the access tools, finding aids and retrieval capabilities of the archive.

* Data mining. The data mining function applies specialized queries or processing functions to the data base to produce new representations of the information content to extend the retrieval capabilities of the data management system.

* Access Reporting. The access reporting function will log usage of all the access functions and provide them to the accounting function.

3.1.7 DISSEMINATION

This entity includes the functions which *receive* data orders from the access function; *monitor* the status of all orders in the system; interact with the storage function to *retrieve* requested data; generate any necessary *ancilliary* information required to accompany the delivery information package; perform any *formatting* requests on the data; make the resulting dissemination information packages available *on -line* or *off-line*; *confirm receipt* of the order and *report* on the successful completion to the administration function.

* Receive data orders. The receive data order function shall accept a data order. A unique order number/identification shall be assigned to each accepted order. For each order accepted, the receive data order function shall add an entry to the pending orders reflecting this accepted order that has not yet been filled. All order information shall be verified, and if any errors or unusual conditions are found, the user shall be notified via an error message sent to the terminal. This function shall provide the user the opportunity to review and correct the information in the order.

* Monitor orders. The monitor orders function shall track every order from inception to delivery confirmation. Operations personnel shall be able to query the pending order file to determine the number and content of each unfilled order. As each order is filled (either by automated or manual means), the order shall be removed from the pending order file. The monitor orders function shall also have the capability to execute a standing order (a standard query and report procedure), based on elapsed time or some other trigger function.

* Retrieve Data. The retrieve function will retrieve the data (from Storage) and metadata (from Data Mgt.) needed to prepare a dissemination information package. The retrieve storage data function shall retrieve data in the storage area and shall logically relocate the data to the staging area for further processing. This function shall accept a request, validate the request using the archive inventory, retrieve the data and place the data in the staging area. The function shall also store the validated selection parameters (field values) from the order in the database. The retrieve metadata function shall retrieve data using relational operators and shall logically place the retrieved data into the staging area for further processing. This function shall accept and validate the commands and store the validated selection parameters (field values) from the command. This function shall then perform the necessary retrieval operations to find and access the requested data.

* Generate ancillary data. This function will generate any additional data or metadata needed to prepare a dissemination information package. The function will determine if

there are any data or software dependencies for a ordered data set and will automatically package these items with the order. It will also access the database to provide user information such as user's name, address, account number, preferred distribution method or media, and other user-oriented information.

* Format data. Format data and metadata into customer specified format. The types of operations which may be carried out include statistical functions, sub-sampling in temporal or spatial dimensions, conversions between different data types or standard output formats, and other specialized processing (e.g. image processing).

* Off line delivery. The off line delivery function shall receive a data order form identifying the contents of a dissemination information package. It shall retrieve the requested data, process the data as required, prepare the distribution media, prepare packing lists, bills of lading, and other shipping records and ship the data. When the order has been completed and shipped, a notice of processed order shall be returned to the administration entity.

* On line delivery. The on line delivery function shall accept a retrieved data set and prepare it for distribution in real time over communication links. The on line delivery function shall identify the intended recipient, the transmission procedures requested, and the data in the staging area to be transmitted.

* Confirm delivery. The confirm delivery function will track a delivery to receipt of data by the customer. The confirm function will e-mail an order confirmation to the user. This function shall notify a user when his order has been executed. The confirmation shall fully identify the order including order number, date of order, date of execution, identity of data requested, and method of distribution.

* Delivery Reporting. The delivery reporting function will collect and record notices and statistics on every delivery to the accounting function. This function will also calculate and record billing information for delivered orders and supply them to the accounting function.

3.1.8 FUNCTION AND SUB-FUNCTION MATRIX

The following matix shows the sub-functions identified under each major functional area and it attempts to align these sub-functions where they have some similar aspects.

Ingest	Archival Storage	Data Management	Administration	Access	Dissemination
Scheduling			Acquisition	Access Control	Receive data orders
Staging	Hierarchy management	Report Request	Configuration Management	Overview (browse)	Monitor orders
Conversion	Physical Migration	Report Generation	*Physical Access Control*	Query	Retrieve Data
Review	Error checking	Update	Planning and scheduling	Retrieve	Generate ancillary data
Extract metadata	Backup		Monitoring	Manipulate	Format data
Create metadata	Duplicate	Database Administration	Customer Service	Display	Off line delivery
Transfer initiation	Transfer receipt		Accounting	Order	On line delivery
			Data Engineering	Advanced Development	Confirm delivery
				Data mining	
Ingest Reporting	Storage Reporting	DBMS Reporting	Administrative Reporting	Access Reporting	Delivery Reporting

3.1.8 Data Flow and Context Diagrams

The flow of data items among the OAIS functional entities is diagrammed in this section. The entities included in these diagrams are the seven that are shown and discussed in Section 3.1.

Figure 3-2 shows the more significant data flows. The flows associated with the Administration and Common Services generally support background activities of the other entities. To avoid complication of Figure 3-2, these background flows are illustrated in the context diagrams of Figures 2-3 and 2-4, respectively.

CCSDS REPORT- REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM

CCSDS 650.0-G-12- April 1997

3.2 INFORMATION MODEL

Section 3.1 of this document presented the Reference Model for Archival Information Systems from a functional decomposition view. This section further describes the pieces of information that are exchanged and managed within the OAIS.

Section 3.2.1 presents a logical model of the information. Section 3.2.2 discusses the details of the issues that make permanent data objects (archive holdings) unique and presents a model for understanding these issues.

3.2.1 LOGICAL MODEL OF INFORMATION IN AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

As discussed in Section 2.1, the primary goal of an OAIS is to preserve information for a designated community over a indefinite period of time. In order to preserve this information an OAIS must store significantly more than the contents of the object it is expected to preserve. Figure 2-1 is an OMT diagram which begins to describe a class structure of the data model for the information contained within the AOIS to preserve a single piece of user information. This section analyzes those classes to fully describe the object classes of data associated with an OAIS. This section uses OMT diagrams to illustrate the concepts discussed in the text.

The Archive Information Package (AIP) contains a content information object and a package of preservation description information objects. The content information contains the object (either physical or digital) to be preserved and the representation information which maps from the physical bit level information into the content concepts addressed by the creator of the digital object. This representation information may be very complex and is discussed in detail in section 3.2.2 of this document. The content object must contain enough representation information for a contemporary member of the user community to

understand the data. Examples of content information are given in section 2.2.1 of this document.

In addition to the content object the AIP must include a set of preservation description information which will allow the understanding of the content objects over an indefinite period of time. This information is typical for all types of archives and has been classified in the context of traditional archives. However, the class definitions must be extended for digital archives. The following definitions are based on the categories discussed in the paper ÒPreserving Digital InformationÓ with examples of each class taken from the discipline of science data archiving. (The definitions are duplicated from section 2.2.2. of this document.)

- Provenance Information: This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This give future users some assurance as to the likely reliability of the Content Information. This information may be thought of as a special case of Context Information described below.

- Reference Information: This information identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Information.

- Context Information: This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created, and how it relates to other Content Information objects existing elsewhere.

- Fixity Information: This information documents the authentication mechanisms,

and it provides any authentication keys used to ensure that the particular Content Information object has not been altered in an undocumented manner.

Ð Catalog Information: This optional information has been extracted from the Content Information to assist in finding the Content Information when searches are made. To avoid the possibility of having to re-extract the information in the future, it is added to the Preservation Description Information set. This information may be thought of as a particular type of Reference Information.

Figure 3-3: Archival Information Package (Detailed View)

Figure 3-3 gives a detailed view of the Archive Information Package. All the "contains " relationships discussed in this section and shown in Figure 3-4 are logical containment relationships. This type of containment relationship may be physical or may be accomplished via a pointer to another object in storage.

The AIPs can be viewed as the "atoms" of information that the archive is tasked to store. These AIPs are then aggregated into an Archive Information Collection(AIC) using criteria determined by the archivist. The AICs are composed of objects and AICs and can reference other AICs of interest (related collections). A logical model of a AIC is shown in Figure 3-4. As in Figure 3-3 all the containment relationships are logical containment and maybe physical or may be accomplished via a pointer to another object in storage. An important feature of AIC as shown in is the fact the a AIC is a complete AIP and there is a preservation description information class which contains further information about the AIC such as why it was created, and related AICs, and fixity information. This is in addition to the preservation description information contained in member AIPs.

Figure 3-4: Archive Information Collection

The AIP and AIC provide the information necessary to enable the long term preservation function of the archive. In addition to preserving information, the OAIS must provide adequate features to allow consumers to locate information of potential interest, analyze that information, and order desired information. This is accomplished through the use of descriptive records (DR) which contain the data that serves as the input to finding aids, visualization aids and ordering aids. Figure 3-5 shows a logical model of an archives full holdings including AIPs, AICs and DRs

Figure 3-5: Archive Holdings Logical View

Finding aids are applications that assist the consumer in locating information of interest. A single AIP may be locatable through a number of finding aids. Visualization aids are applications that allow the user to visualize key features of content object. Ordering aids allow a user to order AIPs of interest. The ordering aids also allow users to specify transformations to be applied to the AIPs prior to disseminations. These transformations can include data object transformations such as subsetting, subsampling or format transformations. The transformations can also involve subsetting the preservation description information in an AIP prior to dissemination. Figure 3-6 provides a more detailed view of the descriptive record (DR). In this figure a class of access aids is introduced as a parent class for finding aids, visualization aids and ordering aids. The information needed for one access aid is called an "associated description". A single DR may contain several associated descriptions depending on the number of different access aids that can locate, visualize, or order the associated AIP. In addition to the associated descriptions, the DR also contains access methods which assist authorized users in retrieving the AIP or AIC described by the DR. In most current archives, only internal archive processes and operations personnel are authorized to use these access methods. However, as technology advances increase the processing power of the archive and the bandwidth betweeen the archive and the user such access methods as Òcontent based queriesÓ and Òdata miningÓ are allowing the archive user direct read only access to the content objects of AIPs

Figure 3-6: Description Record (DR) Logical View

An AIC is created either by creating physical collections of these objects with descriptive records or, more

commonly, by creating "collection descriptive records" which contain relevant metadata about the collection and pointers to contained AIPs or AICs. Currently, this collection metadata is stored in persistent storage such as databases to enable easy, flexible access to the metadata. A copy of these "descriptive records" is included in the preservation description information in the AIP or AIC as "catalog information" to ensure this information is preserved potentially beyond the lifetime of a specific DBMS..

Other than the replacement of the Òdigital objectsÓ with the more structured and complex Òarchive information package,Ó the OAIS logical data model conforms to the Z39.50 Digital Collections Profile [reference 3], which is being widely used as a base standard for digital collection and digital library access. . A more detailed view of the Z39.50 Digital Collections Profile and its relationship to the OAIS information model can be found in Annex C of this document.

Figure 3-7 illustrates the usage of "database management data" within the OAIS. In addition to the "descriptive records" discussed in the previous paragraph, all the information needed for the operation of an archive would be stored in databases as persistent data classes. Some examples of this data are accounting data for customer billing and authorization, policy data, subscription data for repeating requests, and statistical data for generating reports to archive management. These classes are intended as examples rather than an exhaustive list of the data required for archive administration.

Figure 3-7: Database Management Data

3.2.2 Representations of Information

As shown in Figures 2-2 and 3-2, the Content Information object is composed of a Data Object and a Representation Information object. This section further addresses the Representation Information object when the Data Object is specialized as a Digital Object.

The Digital Object, as shown in Figure 2-2, is itself composed of one or more bit sequences. The purpose of the Representation Information object is to convert the bit sequences into more meaningful information. It does this by describing the format, or data structures, which are to be applied to the bit sequences and that in turn result in more meaningful values such as characters, numbers, pixels, arrays, tables, etc. These common computer data types, and aggregations of these data types, are referred to as the Structure Layer information of the Representation Information object. These structures are commonly identified by name or by relative position within the associated bit sequences.

The Representation Information provided by the Structure Layer is seldom sufficient. Even in the case where the Digital Object is interpreted as a sequence of text characters, and described as such in the Structure Layer, the additional information as to which language was being expressed should be provided. This higher layer information is referred to as the Semantic Layer, although in reality each layer provides its own set of semantics. When dealing with scientific data, for example, the information in the Semantic Layer can be quite varied and complex. It will include special meanings associated with all the elements of the Structural Layer, and their inter-relationships. An expansion of the Representation Information object is given in Figure 3-8.

Figure 3-8: Representation Information Object

The Semantic Layer can also be viewed as the "Object" layer. Taking an object oriented view of the combination of the Representation Information and its Digital Object (i.e., Content Information object), queries applied to this object are addressed to the Semantic Layer. The object oriented methods translate these queries into actions on the Strucutre Layer elements, and ultimately to the bit sequences, with results reported back in Semantic Layer terms. Such software methods associated with a Content Information object provide useful services as long as the software executes properly. However for indefinite long-term information preservation,

a full and understandable description of the Representation Information is essential. This can be a particular challenge for the preservation of scientific type data as there are few standards for how to express this type of information and archives need to ensure this information is understandable to the designated consumer communities.

The problem of Content Information migration, which in general can not be avoided over the long term, can be broken into whether the impact is only on the digital object or whether it also affects the Representation Information object. Migrations which only address the movement of bits from one medium to another only replace the Digital Object with a new Digital Object which looks the same at the bit level. There is no impact on the Representation Object except to ensure that it is linked to the correct Digital Object.

Migrations which affect the Representation Information object are much more complex and prone to possible information loss. Those that deal with changes to the Structure Layer information are less of a problem than those that deal with changes to the Semantic Layer information. At the Structure Layer, mappings to a new Strucutre Layer can be said to preserve information if there is an inverse mapping with can reproduce the original Structure Layer information. For example, one can have a mapping from a 16-bit integer to a 32-bit integer, because there is an inverse mapping that works. If more that 16-bits are used in a 32-bit integer, then information will be lost in going to a 16-bit integer.

Changes to the Semantic Layer could involve such things as changing the languge a document is written in; changing the notation used to express relationships among data structures; or applying a new, standard, definition to a data structure. Such changes should, in principle, be reversible but determining that this is true is usually ambiguous.

3.3. HIGH LEVEL DATA FLOWS AND TRANSFORMATIONS

Figure 3-9 presents a high level data flow diagram which depicts the principle data flows involved in OAIS operations. These flows do not include administrative flows such as accounting and billing but concentrate on the flows between an archive and the other entities in its environment (producers and consumers) and internal OAIS flows that involve Information Packages.

Figure 3.9-High Level Data Flows In an OAIS

When an information object is ingested by an archive, various metadata objects are required to assist in the long-term preservation of and access to the information contained in that object. When the information object is stored or migrated, additional metadata objects are created to assist in the preservation and access process. The classes of these metadata objects are discussed in section 3.2 of this document and form the logical information model of the archive. This section discusses the range of potential transformations (both logical and physical) that may occur as an Information Package (IP) passes through the functional areas of the AIS discussed in section 3.1 of this document.

It is important to note that at each point in these data flows, there is a complete Information Package consisting of Content Objects (e.g., images from a spacecraft instrument) and information describing those content object. There are three separate subtypes of IPs(i.e., the Submission Information Package, the Archive Information Package, and a Distribution Information Package) but the information provided in and with the original submission can be preserved in any of these packages.

3.3.1 DATA TRANSFORMATIONS IN THE PRODUCER ENTITY

The data within the data producer entity are private and may be in any format the producer desired. However, when the decision is made to store the data in an OAIS, the scientific investigators who are responsible for the data meet with archivists to negotiate a Submission Agreement.. This agreement spells out the content and format of the Submission Information Package(SIP). The SIP is an Information Package that is provided to the archive by the producer. The SIP consists of the scientific data plus the data that is necessary to assure that those data can be maintained by the archive and that the data can be interpreted and used by scientists who withdraw them from the archive at some time far in the future. These SIPs are periodically transferred to the archive in a Data Delivery Session. The number of data delivery sessions between an archive and a data producer can range from a single session in the transfer a final data product to multiple sessions a day in the case of active archive which store data for experiments which are still in process. The data delivery session can be logically viewed as sets of content data objects and description objects, although physically the description or metadata can be included in the digital objects (i.e., self describing objects) or divided into many separate

descriptive items. In addition to the logical view of data (the SIP), the specification of a data delivery session must also include the mapping of the objects to the media on which they are delivered. This mapping includes the encoding of the object and description and the allocation of logical objects to files.

3.3.2. DATA TRANSFORMATIONS IN THE INGEST FUNCTIONAL AREA

Once the SIP are within the archive, their form and content may change. An archive is not required to retain the information submitted to it in precisely the same format as in the SIP. Indeed, preserving the original information exactly as submitted may not be desirable. For example, the computer medium on which our images are recorded may become obsolete, and the images may need to be copied to a more modern medium. In addition, some types of information such as the Reference ID used to locate the Package within the archive will not be available to the producer and must be input after ingest to the archive.

The ingest process transforms the SIPs received in the data delivery session into a set of AIPs and catalog information which can be stored and accepted by the Storage and Data Management functional entities. The complexity of this ingest process can vary greatly from archive to archive or from producer to producer within an archive. The simplest form of the process involves removing the information and description objects from the producer transfer media and queuing them for storage by the storage and data management functions. In more complex cases, the description objects may have to be extracted from the information objects or input by archive personnel during the ingest function; the encoding of the information objects or their allocation to files may have to be changed; in the most extreme case, the granularity of the information objects may be changed, and the archive must generate new description objects reflecting the newly generated information objects. In addition, the ingest functional entity will classify incoming information objects and determine in what existing collection or collections each object belongs and will create messages to update the appropriate collections after the information objects are stored. During the ingest process, the archive must be highly aware not to unintentionally modify any of the information content in the Producer view. The archive is advised to save the producer media or copy the media into long-term storage as an ultimate reference if needed. It should be recognized that the saving of the producer media will not be permanent because of the issues discussed in Sections 2 and 3.2.3 of this document.

3.3.3. DATA TRANSFORMATIONS BY THE STORAGE AND DATA MANAGEMENT FUNCTIONAL AREAS

The storage and data management functional entities take the AIPs and catalog objects produced by the ingest process and merge it into the permanent archive holdings. The logical model of the ingested data should already be mapped into the logical model of the archives holdings, so the major transformation that occurs in this step is mapping the acquisition session from the ingest physical data model, which will tend to be on staging storage, to the permanent storage of the archive, which could range from database management systems (DBMSs) to hierarchical file management systems (HFMS), or any mixture of the above. The internal view of the archive is the permanent representation of the archived data, so all encodings and mappings must be well documented and understood. The process of transferring the ingest objects is frequently by a software process such as an HFMS driver or a DBMS. In this case, it is the responsibility of the archive to maintain an active copy of the software or careful documentation of the internal formats so the data can be transferred to other systems in the future without loss of information.

3.3.4. DATA FLOWS AND TRANSFORMATIONS IN THE ACCESS FUNCTIONAL AREA

When a Consumer wishes to use the data within the archive, he may use a finding aid to locate information of interest. These finding aids present data consumers with the logical view of the archive so the consumers can decide which information objects they wish to acquire. At a minimum, the access view is the high-level logical view of collections described in section 3.2.1. In most cases, the archive will have spent significant time and effort developing associated description and finding aids such as catalogs that will aid the user in locating information objects or collections of interest. The consumer will establish a Search Session with the access entity. During this Search Session, a Consumer will use the archive finding aids to identify and investigate potential holdings of interest. This searching process tends to be iterative, with a user first identifying broad criteria and then refining the criteria on the basis of previous search results. When the user has identified candidate objects of interest he may use more sophisticated visualization aids such as browse image viewers or animation to further refine his result set.

Once the Consumer identifies the archive holdings he wishes to acquire, he must issue an order request to the archive to acquire the data. This order can also specify any transformations the Consumer wishes applied to the AIPs in creating the Dissemintation Information Package (DIP). The request triggers the dissemination process, which is discussed in section 3.1.7 and 3.3.5.

Archives and external organizations may provide additional associated descriptions and finding aids that allow alternative access paths to the information objects of interest. As data mining technologies become more mature, it is likely that researchers will develop new and fundamentally different access patterns to information objects. It is important that an archive's access and internal data models are sufficiently flexible to incorporate these new descriptions so the general user community can benefit from the research efforts. A good example of this type of new associated description are a phenomenology database in Earth Observation, which allows users

to obtain data for a desired event such as a hurricane, or volcano eruption from many instruments with a single query. It is important to note that such finding aids may become obsolete unless the data they require are preserved as parts of the AIPs they access.

3.3.5. DATA FLOWS AND TRANSFORMATIONS IN THE DISSEMINATION PROCESS

The dissemination process serves the data consumer in roles very similar to the way the ingest process serves the data producer. Through interactions with the access functional area, the data consumer produces a logical view of the desired information objects and descriptions to be include included in the Dissemination Information Package. At this point, the consumer issues an order request for this DIP that triggers the dissemination process, which negotiates a request agreement with the customer in which the physical details of the Data Dissemination Session such as media type and object format are specified.

The Dissemination functional area then contacts the Storage and Data Management functional areas and requests the AIPs and catalog data necessary to populate the DIP requested by the consumer. The Storage and Data Management functional areas create copies of the requested objects in staging storage.

The Dissemination process transforms this set of AIPs and catalog objects into a DIP and stores that DIP onto physical distribution (either physical or communications) media to be delivered to the data consumer in a Data Dissemination Session. The complexity of this transformation process can differ greatly on the basis of the level of processing services offered by the archive and requested by the data consumer in his order. In the simplest case, the DIP contains duplicates of the AIPs and catalog objects of interest from storage and data management function. In more complex cases, the description objects may have to be extracted from the information objects or inserted into self-describing information objects, and the encoding of the information objects or their allocation to physical files may have to be changed. In the most extreme case, when the archive supports subsetting services, the granularity of the of the information objects may be changed, and the Dissemination process may generate new description objects reflecting the newly generated information objects.

4 MIGRATION PERSPECTIVES

The fast-changing nature of the computer industry and the ephemeral nature of electronic data storage media are at odds with the key purpose of an archive: to preserve information over a long period of time. No matter how well an archive maintains its current holdings, it is likely that over time much of the stored information will need to migrate to different media or to a different hardware or software environment to remain accessible. Each archive must have a general plan for data migration. Specific requirements for, or limitations on, the migration of individual datasets should also be spelled out in the data submission plans.

The general rule for data migration is that information content should be preserved although the data itself may change. The information content of a copied dataset is "equivalent" to the original if there is a known transformation that can generate the original information from the copy. Usually this means that data should not be deleted during data migration unless those data are truly redundant. Data may be modified or added as long as the information content is preserved. For example, ASCII character-coded data may be migrated to a new character set if the new code is a superset of ASCII. As another example, a lossless compression could be applied to data during migration since the compression is reversible. In practice, full equivalence may or may not be necessary or even achievable when migrating specific datasets. In cases where full equivalence is not provided, then appropriate metadata should be attached to a dataset to describe how the copy differs from the original data.

Whenever archived data are moved or modified in the ways described below, the Provenance Information should be updated to document the migration process. The associated Representation Information will often also need to be updated. For example, migration of data to a different kind of media will typically require the updating of Representation Information that maps the logical structure of an information package to physical volumes. A migration plan should include the mechanisms for changing Representation Information and for verifying the information after migration is complete. Here again the equivalence principle holds: if the new Representation Information is correct and self-consistent, and if it can be mapped back to the original Representation Information, then the information content has been preserved. Updates to Fixity Information may be needed if the migrated data are encrypted or decrypted. Catalog metadata may also need to be updated as a result of migration.

4.1 MEDIA MIGRATION

Probably the most common sort of data migration is to a new storage medium. Today's data storage media can typically be kept for a few years to a few decades before the probability of irreversible loss of data becomes too

high to ignore. An archive should have a policy that dictates how each type of media is to be stored, monitored, and ultimately replaced. Media manufacturers and industry associations can provide much of the information needed to formulate this policy.

If an original volume is copied to the same medium with no changes or additions to the data then there is a simple standard for preservation of information: if the copied data match the original data bit for bit, then information is preserved and compatibility with any hardware and software that can read the original data is assured. Not infrequently a volume will be copied to either an upgraded version of the same medium (for example, a higher-density tape) or to a different medium (for example, from tape to optical disk). In such cases, two issues need to be addressed: when is the copied information equivalent to the original; and what effects will the migration have in terms of the hardware and software used to access the data? If the original bit stream can be reproduced from the copy, then the copy and the original are equivalent across media. If software can reproduce from the copied data the results achieved on the original data then the copy and the original are equivalent across hardware/software environments.

4.2 LOGICAL STRUCTURE

When migration occurs, the organization of information into logical volumes, directories, files and records may change. The logical structure of a copied dataset is equivalent to the original structure if the old structure can be reproduced from the new. This means, for example, that directory paths and file names can change, provided that there is a one-to-one mapping between the old and new names. If new information is introduced during migration, it is better to keep the new material separate from the original material (for example, new material should be added as new files rather than modifying original files).

4.3 DATA OBJECTS

The information conveyed by a specific data object is preserved if the original object can be precisely and completely reconstructed from the copy. This allows for superficial changes to a data object -- for example, a change in byte ordering -- but this also allows for more complex transformations.

For primitive data objects, a general set of rules can be given:

- For character-coded data, the character set may change as long as the new set can represent all of

the old characters. For example, updating a dataset to use the new Unicode standard would be admissible for data currently coded in ASCII.

Ð For numeric data, two factors come into play: magnitude and precision. For integer numeric values, only the magnitude is important; integer numeric data can be changed to a different integer representation if the new format can represent the old data's maximum positive and negative values. For fixed-point real numbers, both magnitude and precision requirements must be addressed. Floating-point real numbers present the most difficult case, since both the magnitude of the exponent and mantissa components must be considered. More importantly, precision may change under different floating number formats. It should be noted that the common solution of storing real numbers as character-coded values is not guaranteed to preserve information since the conversions from binary format to character format and back may differ across hardware/software environments. One solution is to always use hardware that adheres to a standard -- like the IEEE-488 standard -- for the representation of numeric data.

Ð For complex data objects -- composed of a set of primitive objects -- the equivalence principle allows for many kinds of updates, including changes in underlying representation (using floating point rather than integer numbers), change in storage order (column major versus row major), and even lossless compression of the data. Some migrations, however, may not preserve information in the strictest sense. Examples are lossy data compression and remapping of data to a different map grid. The issue of information preservation then becomes a subjective matter to be decided by the information producers, archivists, and information users: if modifying the data does not alter significantly the results that users arrive at when they examine or employ those data, then the meaningful information has been effectively preserved. If there is doubt about whether or not significant information will be lost during a migration, it is recommended that the original information be disseminated out of the archive and a new information package generated, ingested back into the archive, and then maintained along with the original information.

5 ARCHIVE CLASSIFICATIONS

The Reference Model for an Open Archival Information System is meant to cover a wide range of possible implementations and therefore a variety of archives are expected to use the standards which will be based on the reference model. Given such a diversity, it is useful to have a uniform manner for describing or 'classifying' archives; both as a means of comparing two archive systems and of describing one's archive to one's management. While the state of describing archives has not yet progressed sufficiently far to rate them on an absolute scale, the points enumerated and discussed below do allow a context for such discussions.

1. Acknowledged degree of permanence:

Classification criteria:
a) Temporary archive: archive has a defined lifetime and is not expected to exist beyond that time.
b) Permanent archive: archive is defined as permanent, or the date at which it may be terminated may be extended after additional review by management.

2. Digital Information preservation level:

Classification criteria:
a)ÊBit Preserving: archive is responsible for preserving collections of bits
b) Information Preserving: archive is responsible for preserving bits as well as what those bits mean.
(this does not address the physical media/samples issue)

3. Degree of opaqueness of AIP

The traditional view of archives is that the objects are opaque to the access function of the archive (ref. Digital Collections Profile). In other words, the access function can only process queries based on meta-data (catalog entries) which have been explicitly stored for each object. This does not mean that information could not have been extracted from the objects as they were ingested (or at a later time) and that this information was attached to the object as meta-data. This traditional barrier is beginning to breakdown in some archives (or digital libraries) and "data-mining" techniques may be available in some cases.

Classification criteria:

a) Objects are completely opaque to the access function.
b) Some information within an object is directly available to the access function.
c) Objects are available to the access function.

4. Dissemination methods

Traditionally, archives have delivered ordered products on media (paper ,magnetic tapes, CD-ROM, FTP sessions). With the growth of communications bandwidth and Internet technologies there is a trend towards making archive access and dissemination seem like a single interactive process. This view can change the underlying architecture of the archive and severely restrict the maximum size of DIPs.

Classification criteria:
a) Non-electronically readable media (paper, microfilm, etc...).
b) Electronically readable media.
c) Electronic transmission (internet, modem, etc...)

5. Active vs Final Archive

The traditional archive provides storage, access and dissemination of unchanging artifacts. However, many current scientific ÒarchivesÓ combine the functionality of data processing, data access and data dissemination for science products that are being created for the first time. These Òactive archivesÓ must deal with all the issues of traditional archives but have several unique problems such as the ingest of new ÒversionsÓ of currently archived objects, changing metadata for existing AIPs, and continuous ingest processing demands. Also the relative priority of the ingest, access, preservation and dissemination functions may differ in active archives vs traditional archives.

Classification criteria:
a) Active archive
b) Final archive

6. Diversity of Collection

Archives may support any level of heterogeneity of the subject matter in their collections. The level of heterogeneity of the collections will affect several factors including:

-Degree of quality control
-Degree of user support
-Who oversee quality control

- Degree of discipline expertise (or specific data expertise)
-Number of SIP formats accepted
-Sophistication of finding aids

Classification criteria:
a) Project
b) Discipline
c) General

7. Institutional vs Non-Institutional archive:

Ultimately, an archive serves the interests of the organization which gave it its charter and (one hopes) provides in some manner for its funding. However, the source of objects and/or the reason for storing the objects may or may not be directly related back to this organization. Often the community supplying or retrieving the objects is larger than the chartering organization or the chartering organization is only a part or representative of the community. If the archive is an integral part of the organization used primarily for preserving its own records, than it would best be described as an Institutional Archive.

Examples of Institutional archives: NARA, State government, Coca-Cola.

Classification criteria:
a) Institutional
b) Non-institutional

8. Archival storage types

Classification criteria:
a) Physical: includes physical samples, hard copy, film, etc.
b)ÊDigital: Information archived is in digital forms
c) Both: Some information archived is in digital and some is in phyiscal forms

9. Distributed vs Centralized

This archive reference model does not specify whether the functional areas are physically centralized or distributed. However the current model assumes a single administration and management function which sets the operational policies for all functional areas. The use of the OAIS model with federated archives will drive additional administration and data flow considerations.

Classification criteria:

a) Centralized

b) Physically distributed, centralized administration/management

c) Physically distributed, federated administation/management

6 ILLUSTRATIVE SCENARIO

This scenario describes the flow of information into and out of a hypothetical archive of space science data. Data from spacecraft Ñ monitoring, for example, the EarthÕs climate Ñ may have as much value to future scientists as to the scientists of today; therefore, the data from spacecraft are often archived with the intent to preserve them for decades, even centuries. Spacecraft data are often stored in an *active archive* Ñ an archive where data is flowing into the archive over an extended period, rather than as a single submission. In this scenario, the latest data from the spacecraft is transferred to the archive every day for safe keeping. But the process begins months, perhaps even years, in advance of the launch of the spacecraft. At that time, the scientific investigators who are responsible for the data meet with archivists to negotiate a *Submission Agreement.*. This agreement spells out the content and format of the *Submission Information Packages* that will be provided by the producer to the archive. Each SIP consists of *content information Ñ* the scientific measurements that are the core of the submission Ñ plus *preservation description information* that is necessary to assure that the content information can be maintained by the archive and that the data can be interpreted and used by scientists who withdraw them from the archive at some time far in the future.

Our spacecraft carries a scientific instrument that takes images of the Earth, so the Content Information in our submission packages are a set of digital data objects, with each object a single image taken by the instrument. Along with these images, the scientists agree to supply *representation information* Ñ computer-readable information describing the format of the images. For example, the representation information may specify that each image consists of 1000 scan lines, with each scan line containing 800 pixels, and with each pixel represented by an unsigned 16-bit integer value. The notation for specifying this kind of information is often human-readable as well machine readable, in which case it is called a Data Description Language (DDL). Using a hypothetical data description language, we may describe the characteristics of our images as follows:

```
Object EARTH_IMAGE is
   SCAN_LINES = 1000;
   PIXELS_PER_SCAN_LINE = 800;
      PIXEL_TYPE = UNSIGNED_INTEGER;
      PIXEL_SIZE_IN_BITS = 16.
```

This DDL specification applies to all of our images. How does one know what each term in the definition (for example, PIXEL_TYPE) means? Often the terms used to describe data are themselves formally described in an information repository know as a data dictionary. For example, the data dictionary entry for PIXEL_TYPE may indicate all of the possible types of pixels (integer, floating point, etc.) that can be described using a particular DDL. Since every image returned by our spacecraft will have the characteristics cited above, the DDL formalism for our images, along with the data dictionary describing the terms we use in the DDL, need only be submitted to the archive once, rather than with each submission of new data to the archive.

Throughout the spacecraftÕs mission, a daily cycle is followed, as shown in Figure X-1: raw data is transferred from the spacecraft to a ground station; the data are processed; and the resultant images transferred into the archive. Each SIP comprises all of the images transferred to the archive during a specific day, along with the following preservation description information:

Ð Provenance information Ñ This information describes how the data were processed or handled before being inserted into the archive. It includes a processing history describing briefly each process that was applied in the preparation of the image. For example, our images have been calibrated to convert the unique output of the instrument into a physical quantity (like brightness temperature) that can be compared with other scientific data. The calibration that was used in this process is included in the archive as provenance information. Some provenance information

may be submitted infrequently (our calibration need only be submitted to the archive whenever it changes); other provenance information may have to be included within each SIP submitted to the archive.

Đ Context information Ñ Information that relates our images to other data sets. For example, context information may be provided to indicate where to find the data file that describes the spacecraftÕs position and orientation in space as a function of time. Having access to this information allows us to determine the latitude and longitude of every image.

Đ Catalog metadata Ñ Information about the submitted images that is intended to make it easier for users to locate images that are of specific interest to them.

As the information cited above is ingested into the archive, the form and content of the archive package may change. An archive is not required to retain the information submitted to it in precisely the same format as in the SIP. For example, in our case the images may be subjected to lossless data compression to reduce storage space on disk and tape. Also some of the representation information will not be archived with the images; instead they are transferred into an online database so that the images can be located and retrieved by archivists and external users.

After a while, the computer medium on which our images are recorded may degrade or become obsolete, and the images may need to be copied to another medium. This process is called data migration. The fundamental rule of data migration is that the information content must not be diminished. For example, our images with 16-bit pixels cannot be converted to have only twelve bits per pixel, because information would then be lost. The archiveÕs operating policies, along with the Submission Agreement, define what constitutes Òpreservation of informationÓ for a particular data set.

Figure X-1. Spacecraft Data Archiving Scenario

When a scientist wishes to use the data within the archive, he may use a finding aid to locate information of interest. For example, a scientist may indicate that she wants all images taken over the state of Colorado during the month of October. The finding aid would search through the catalog metadata to identify the set of images within the archive that meet this condition. The scientist can then place an order for these images. When the order is processed, the images are copied from data storage and formatted for transfer to the scientist. Here again the format of the images or the associated information may change, but the information content must be preserved. Along with the images will come representation data to help the scientists and the software they are using to interpret the images. Preservation Description Information is also provided to let the scientists know the pedigree of the data. This would include, for example, information on the processing performed prior to the submission of images to the archive, along with any changes made to the original images while they have resided within the archive. Once the entire Dissemination Information Package of information is prepared, it is transferred to the

scientist over a wide area network.

At each point in this scenario, there is a complete Archive Information Package consisting of Content Information Ñ images from the spacecraft instrument and associated representation information Ñ and information describing the preservation of the content information. There are three separate packages Ñ the Submission Information Package, the Archive Package, and a Dissemination Information Package Ñ but the information provided in and with the original submission is preserved. Part of the representation and preservation information may be submitted separately from the main content information (images), but when the images are stored in the archive the representation information is already available, thus completing the package. The concept of digital archives as represented by the reference model in this paper, encompasses traditional archives Ñwith bulk deliveries of data that are no longer needed or wanted by the organization that generated them Ñ as well as active archives where data are delivered regularly and where the data may be frequently accessed by users.

ANNEX A: SCENARIOS OF EXISTING ARCHIVES

A.1 PLANETARY DATA SYSTEM ARCHIVE

I. Domain and Customers.

The Planetary Data System is chartered to provide data archiving services, data access and expert help to the NASA-funded planetary science community. The PDS is a distributed system with a Central node at the Jet Propulsion Laboratory and discipline nodes (imaging, geosciences, atmospheres, planetary plasma interactions, small bodies, rings) located at universities around the country. The early focus has been on restoring historical mission data and has produced several hundred CD-ROM volumes containing about 80 per cent of the important planetary data archives. There has been an increased emphasis on providing access to the general public for educational outreach over the past several years.

Data Producers

Planetary data sets originate with NASA flight project data management and science teams (new data, some restorations), individual scientists (newly processed or value-added data) or via the PDS discipline nodes (restorations and value-added data). At least 50 per cent of the PDS resources have been devoted to restorations over the past seven years, with several more years of work needed to capture all historical data.

II. Ingest Process and Ingest Interface

The PDS has developed a very formal interface with the major data producers (flight projects). This interface is documented in the Data Preparation Workbook and involves substantial interaction between node personnel, data engineers and project representatives. A Project Data Management Plan, signed by the PDS project manager provides the basic

project data description and agreement to deliver to PDS. Since about 1993 all NASA announcements for Planetary investigations or analysis require that all data generated be delivered to PDS in conformance with PDS standards.

Submission Agreements

¥ Projects provide a Project Data Management Plan. Sometimes a more specific document, the Archive and Transfer Plan supplements the PDMP, providing extended product documentation and a schedule of deliveries.

¥ Individual scientists can propose to be "data nodes" and receive funds from a PDS discipline node for preparing restored or value- added data sets for inclusion in the archive. There is no formal submission agreement for data nodes.

¥ The PDS discipline nodes each maintain a list of outstanding restorations. These are worked-off based on their priority within discipline. At some point this list will be completed and only new project or data node data sets will be ingested into PDS. There is no formal agreement associated with discipline data restorations.

Each data set that is identified for ingest in PDS is assigned to a Central node data engineer. It is the responsibility of the data engineer to see that all archiving steps are completed. The archiving steps are called out in the PDS Data Preparation Workbook.

Describe a typical data delivery session. Typically a delivery session will consist of a single data set contained on one or more volumes of CD-ROM or CD-Recordable media. A data set is defined within PDS to be a group of homogenous data granules at the same data level (raw, decalibrated, reduced) which differ only in time of acquisition and major category of target body. For example, the images of Jupiter taken by both Voyager spacecraft comprise a single data set. The standard process includes up-front negotiations between PDS and the provider; the production of test products which are evaluated in the

peer review; revised final test products which are validated by the data engineering staff at the central node; approval and production of CD-ROM volumes; distribution by the appropriated discipline node or the central node; entry of the data set into the PDS central catalog; and entry of the data set into the NSSDC ordering system.

Transformation Process

In most cases the original data formats are maintained when data is brought into PDS. This allows existing software tools to continue to be used with the data. Much of the data preparation involves carefully documenting the data format and preparing metadata (granule labels, index files and catalog templates).

Validation

Validation is generally performed as part of the peer review of a product or by using validation tools. In some cases (for example, Magellan), the project develops its own internal validation process. The main validation tool of the PDS is the Volume Verifier. This program is run by the Central Node data engineers on each product delivered from a project or a data restoration. It validates the format and content of all product labels, and validates data files using checksums.

Security

The only area where any special security issues exist involves the receipt of proprietary data. Some projects have one-year proprietary periods before data in released to the science community. The PDS policy is to avoid receipt of any proprietary data sets during the proprietary period.

III. Internal Forms

The PDS has developed standards for documenting data sets (templates) and individual data products (PDS labels) using a keyword=value label system called the Object Description

Language (ODL). Recommendations are also provided for volume organization and data product formatting to optimize the utility of resulting data products.

The PDS standards are specified in the PDS Standards Document. Standard documentation requirements include templates describing the data set, instrument, mission, etc. These templates are included on data volumes and also entered in the PDS high-level catalog. Standard terminology is maintained in the Planetary Science Data Dictionary, which is jointly maintained by the PDS and the multi- mission ground data system. The metadata values for new data products are carefully compared with the PSDD and existing values used wherever possible. Additions are made to the PSDD to add new standard values to accomodate new datasets and when justified new keywords are added to the PSDD. Data products can have specialized metadata values which are not cataloged in the PSDD.

The PDS product labelling system is flexible enough to allow nearly any data structure to be described. Labels can be attatched to the beginning of the data file or detached in a stand-alone text file which points to the data file. In some cases a single label file is used to describe multiple data files. Detached labels can be used to describe data stored in other formats (FITS or HDF, for example). In cases where complicated raw telemetry formats are stored the Software Interface Specification (SIS) for the product is included in lieu of descriptive labels.

Archive Volume Components.

An archive quality dataset is required to contain the following components.

AAREADME.TXT - Text summary of data contents.
VOLDESC.SFD - Standard volume label.
VOLINFO.TXT - Text description of data contents.
CATALOG - DATASET.CAT, MISSION.CAT, INST.CAT
INDEX - ASCII index for each granule on the volume.

SOFTWARE - Software needed to interpret/display the data.
CALIB - Calibration data sets.
BROWSE - Browse products for this volume.

Peer Review.

All restoration and data node produced data sets are required to undergo a peer review before acceptance as archive products. Products produced by flight projects do not go through a formal peer review process. In general there is ongoing negotiation between the data engineer or the discipline node staff and the data producer. The peer review team consists of a number of scientists familiar with the data set, the discipline node leader and one or more data engineers. All product documentation and sample products and software are supplied to the peer review group for evaluation. The peer review group determines the adequacy of documentation and quality of the data products and either approves the product or provides a set of liens which must be fixed prior to approval. The PDS nodes and data engineers have access to a Volume Verifier tool which aids in validating the quality of an archive volume. The volume verifier checks internal checksums, verifies that the index contains entries for all data products and validates the volume templates as well as the descriptive keywords supplied for each product.

Delivery Media.

Discipline restorations and data node products are recorded on CD- ROM or CD-recordable media as a standard practice. Flight projects are urged to provide archive quality products on CD media but may not be able to due to funding constraints. Products delivered to PDS on magnetic tape media are assigned to the PDS restoration queue. It is the goal of PDS to convert all data sets to CD-ROM or CD-recordable media which is replicated at a separate geographic facility. This separate facility is generally the National Space Science Data Center (NSSDC) at Goddard Space Flight Center.

IV. Access

Nearly all access to PDS data sets is via the CD-ROM volumes which are distributed to the entire research community. Large discipline node data collections including a substantial volume of CD-ROM data are accessible via the internet. Several of the discipline nodes have developed on-line retrieval systems customized to meet the needs of their discipline scientists.

Finding Aids

The Pilot PDS devoted substantial resources to designing a central catalog system and distributed query and processing capabilities at the discipline nodes. These efforts were largely dropped as the Planetary Data System focused on data restoration rather than data access. In general, most of the user community already had home grown tools for data analysis and were most concerned with getting access to the data sets. The growth of the user community due to internet and increased usage of CD-ROM readers has spurred to prototype a more consistent finding aid. The PDS Navigator has been developed for selecting images from the Clementine mission. It includes three components, a forms-based traditional database retrieval capability, an image-based retrieval and a text-based retrieval.

Security

The high-level PDS catalog can be accessed via a group account. Most of the data access services at the discipline nodes require the user to obtain a valid account on the node computer.

Customer Service/Support

The order function of the PDS is distributed. Data inventories are kept at NSSDC, the PDS central node and at each discipline node. In general each site serves a special group of users:

- discipline node - members of the NASA funded discipline
- central node - other NASA scientists and engineers, other agencies
- NSSDC - other scientists, agencies, public and foreign users.

The PDS Operator at the central node handles requests for PDS documentation or standard data products. The discipline nodes handle data requests from within their discipline and also provide expert help in the utilization and interpretation of the data. Access to tools is also provided.

V Dissemination

The vast majority of data dissemination is done via CD-ROM disc. Several hundred copies of over 500 titles have been distributed to date.

Subscriptions

Nearly all PDS distribution is done via subscriptions or standing distribution lists. It is the responsibility of each discipline node to maintain a distribution list for its discipline scientists. This list determines the order amounts for most CD-ROM titles. The central node maintains a distribution list for engineering and management personnel and for other external reciepients (reciprocal distribution, software developers).

Media/network use

Nearly all final products are delivered to the user community on CD- ROM. Archival products that need not be widely distributed are stored on CD-Recordable media, with a duplicate copy provided to the NSSDC. Most PDS data is available for downloading via anonymous ftp connection to a large CD-ROM jukeboxes at the central node and the imaging node.

Data Manipulation

Each discipline has a suite of government developed analysis tools which can be applied to the discipline data sets. These software packages are available for UNIX workstations or VAX VMS platforms. Several nodes provide the user a menu of processing functions that can be performed on selected data and will carry out requested processing and provide the results electronically or via media. The most widely used commercial tool is IDL.

Pricing policy

The PDS distributes data to legitimate NASA researchers for no charge. There are no charges for on-line computer usage or data processing to NASA researchers. The NSSDC distributes CD-ROMs for $10 per volume.

Security

All PDS data sets are certified GTDA by the Department of Commerce and are distributable worldwide.

VI Special Characteristics

PDS has invested a substantial engineering effort in its common data dictionary, data standards and procedures for preparing archival quality data sets. By having these standards in place the PDS is able to demand better quality data sets of its data providers.

Each discipline has a suite of government developed analysis tools which can be applied to the discipline data sets. These software packages are available for UNIX workstations or VAX VMS platforms. Several nodes provide the user a menu of processing functions that can be performed on selected data and will carry out requested processing and provide the results electronically or via media. The most widely used commercial tool is IDL.

Pricing policy

The PDS distributes data to legitimate NASA researchers for no charge. There are no charges for on-line computer usage or data processing to NASA researchers. The NSSDC distributes CD-ROMs for $10 per volume.

Security

All PDS data sets are certified GTDA by the Department of Commerce and are distributable worldwide.

VI Special Characteristics

PDS has invested a substantial engineering effort in its common data dictionary, data standards and procedures for preparing archival quality data sets. By having these standards in place the PDS is able to demand better quality data sets of its data providers.

A.2 National Archives and Records Administration's Center for Electronic Records

I. DOMAIN

Domain and Customers

The Center for Electronic Records is the organization within the National Archives and Records Administration (NARA) that appraises, accessions, preserves, and provides access to federal records in a format designed for computer processing. Customers for this data are as diverse as the electronic records they seek to access and range from individuals seeking to assert their rights to other government agencies to academic researchers, private consultants, media personnel, and a wide variety of other users.

Data Producers

Originally these records are created or received by agencies of the federal government. They may concern virtually any area or subject in which the government is involved. They may come from any type of computer application such as data processing, word processing, computer modeling, or geographic information systems. They can include records made directly by government employees or indirectly through government grants and contracts.

Special Features

The most noted special feature is the diversity of the collection of more than 23,000 data sets from more than 100 bureaus, departments, and other components of executive branch agencies and their contractors and from the Congress, the Courts, the Executive Office of the President, and numerous Presidential commissions. A few of the data files were originally created as early as World War II. An even smaller number contain information from the nineteenth century that has been converted to an electronic format. Most of the holdings, however, have been created since the 1960s. The major types of holdings and subject areas include agricultural data, attitudinal data, demographic data, economic and financial statistics, education data, environmental data, health and social services data, international data, and military data.

Scientific and technological data already transferred to the Center include the National Register of Scientific and Technical Personnel; the National Engineers Register; the 1971 Survey of Scientists and Engineers; major portions of the National Ocean SurveyÕs Nautical Chart Data Base; numerous Environmental Protection Agency series relating to pesticide use, hazardous wastes, and pollution abatement; the Nuclear Regulatory CommissionÕs Radiation Exposure Information Reporting System; biometric data sets and epiemiological studies such as the National Collaborative Perinatal Project from the National Institutes of Health, the Centers for Disease Controland data, and the National Center for Health Statistics; and text from presidential commissions on Three Mile Island, coal, and the Space Shuttle Challenger Accident. While the CenterÕs scientific and medical holdings are rich and varied they do not fully reflect the extent and diversity of Federal activity in this area.

II. INGEST

The ingest process begins with records managers and records creators in federal agencies inventorying all electronic records and determining how long to retain the records for current agency business. The next step in the process is for the agency and NARA to develop a *Request for Records Disposition Authority*, Standard Form 115 (SF 115). Here information on the content, retention and disposition, and the availability and extent of documentation and related reports is listed in the context of the creating program and current agency business needs for the information. Records with continuing value are listed as permanent and the timing and frequency of their transfer to NARA is established. The SF 115 is submitted to NARA for its review and appraisal. The Center for Electronic Records appraises electronic records items on all SF 115s. Identifying permanently valuable electronic records for retention by NARAÕs Center for Electronic Records involves cooperation between NARA and the various federal agencies. Through the process of scheduling and appraisal, the Center identifies and selects the electronic records it judges to have enduring value. The Center evaluates electronic records in terms of their evidential, legal, and informational value and their long-term research potential. Some of the factors in this appraisal evaluation include the value of the data in the context of past, present, and probable future research trends within the context of the dataÕs origin and current use and its impact on federal programs and policy. Administrative and legal value, as well as the potential for linkage with other data, may bear on the decision. Unaggregated microlevel data sometimes has the greatest potential for future secondary analysis. Once the Center determines the records have enduring value, it then determines whether the records should be preserved in electronic format.

Submission Agreements

The actual submission agreement between NARA and the agency that creates or receives the records is a *Request to Transfer, Approval, and Receipt of Records to the National Archives of the United States*, Standard Form 258 (SF 258). It transfers physical and legal custody of the electronic records from the federal agency to NARA. This agreement is the end product of the ingest process described above. The SF 258 also contains any restrictions on access to the data which conform with exemptions listed in the Freedom of Information Act. The Center will uphold all legitimate restrictions on access. The Center also will work with the creating agency to determine if any Òdisclosure-freeÓ version of the data can be produced for researchers.

Typical Delivery Session

This inventorying, scheduling, and appraisal process specifies the electronic records and related metadata and documentation to be transferred and establishes the transfer times and frequencies. Specific instructions for how the records are to be organized and when they should be transferred are established in the *Code of Federal Regulations* (36 CFR 1228.188). All electronic records should be transferred on either open reel magnetic tape, tape cartridges, or CD-ROM. The CFR sets the specific technical requirements in terms of format, block size, and extraneous characters. While the current regulations also require that all electronic records should be transferred in a software-independent format, NARA staff recognize that the research potential and utility of some electronic records would be significantly reduced if they were transferred in such a format. In such cases NARA works with the creators to determine the best mode of transfer.

What are the Information Objects that are Delivered? Agencies typically will transfer a series consisting of one or more data sets with the related documentation which minimally should include the record layout and codes, methodology statements, technical information about the data including number of records and size, and desirably also includes associated analyses and reports. Increasingly agency-created metadata also is included. The majority of electronic records come as flat files of data; increasingly, however, text files and output from data base management systems, and geographic information systems also are transferred.

What are Collections? NARA organizes all records on the basis of Provenance and Original Order. Provenance maintains the identity of a body of records and preserves as much information as possible about its origins and custodial history. Within NARA this is accomplished through the use of Record Groups which reflect the structure of the federal government and subgroups and sub-subgroups which place the collections within the creating unitÕs place within its agency. Original order argues for the maintenance of the original order of a body of records. It reveals the creatorÕs organization and use of the records and can provide additional information to secondary users. For electronic records, Òoriginal orderÓ is expressed in the logical structure of files and databases and in the indexing which the creator used. Within NARA the basic unit for arrangement and description is the series which can include a number of related data files.

What Descriptive Information is Provided? The extent and quality of the descriptive information provided by the creating agency varies from quite sketchy to extremely detailed. NARA staff attempt to flesh out the agency-created information with series level descriptions, title list entries, abstracts, and documentation packages and to provide the descriptive information in a variety of formats to reach different research clienteles.

What sorts of Validation Objects are Provided? Agencies are required to transfer documentation adequate to access, process, and interpret electronic records. For formatted data files the documentation must include a record layout with appropriate field definitions and codes. It frequently also includes methodology statements, input documents, data entry instructions, processing directions, sample outputs, reports and analyses of the information and system manuals.

What Transformation Processes are Performed Prior to Storage

What Metadata is Created? The most extensive metadata product created by NARA is the documentation package. In the Introduction, Center staff discuss the origin, creation, and administrative uses of the records, list related records that are or will be available, and discuss characteristics of the records that could cause problems for researchers based on initial validation processes. The documentation also includes sample printouts of the data and tables and reports related to computer validation of the data. NARA also capyures metadata on record layouts, domains, ranges, and links between files in a metadata database as a byproduct of the automated validation process. Other metadata created by Center staff include series descriptions, formatted abstracts, title line entries, and collective descriptions which place the records in a broader context. The Center anticipates that increasingly metadata created by the originating agency will be part of the information transferred to NARA.

What Validation is Performed? The CenterÕs initial accessioning procedures include creating a new master and backup copy of each submission on new certified media to ensure the best physical media for long-term storage. At this time Center staff perform automated comparisons of the data contents with the record layout and codes, and of the physical structure including the number of records, blocks, and bytes. Staff also perfect the documentation package to facilitate secondary use of the data.

Security

All data are maintained off-line with researcher access only to copies of the data. The master and backup copies are maintained in separate secure stacks at two different physical locations. Data which require additional security measures, for example Census data subject to restrictions imposed under Title 13 of the *United States Code* and national security classified information restricted under Executive Order, are afforded the appropriate level of protection. The Center is moving to provide enhanced access to selected data onsite by providing reference copies on a wider variety of media and by providing a broader range of services and output products. This may include use of vendors who can provide enhanced access to the holdings utilizing Òvalue-addedÓ services.

III. INTERNAL FORMS

How do you Store your Data? All master and backup copies are stored on newly certified class 3480 magnetic tape cartridges. Some of the holdings have not yet been migrated from nine-track, 6250 bpi open-reel magnetic tape. Data are received and stored temporarily on other media including diskettes, 4mm, 8mm, CD-ROM, and various removable hard drives, although not all of these media conform with regulatory requirements.

Migration (Data). Based on recommendations from the media manufacturers, the National Media Laboratory, the National Institute of Standards and Technology, and various standards organizations, the Center has been migrating its data to new class 3480 magnetic tape cartridge when each media unit is ten-years old.

Migration (Metadata). Metadata has been stored in a variety of formats depending on the original format transferred with the data. Traditionally most metadata existed in textual format. The metadata captured in the validation process is maintained in a relational database. There are no current plans for migrating from this format, although the metadata can be exported in flat file format. The Center has been encouraging data creators to create and transfer metadata in electronic form. Within the next fiscal year the Center hopes to begin scanning and digitally converting metadata so it can be preserved and provided in an electronic format along with the data.

Migration (Format). The *Code of Federal Regulations* requires data creators to transfer all

data in ASCII or EBCDIC with all extraneous characters removed from the data except record length indicators or tape marks and blocked at no higher than 32,760 bytes per block for open-reel and 37,871 bytes for class 3480 magnetic tape cartridge. When CD-ROM is used they must conform to ISO 9660 standard and the files must be discrete files containing only the permanent files. Additional software files and temporary files may be included on the CD-ROM. The CFR also requires all electronic records to be transferred in a software-independent format. The Center works with data creators who cannot meet those requirements to determine the most appropriate transfer and storage formats.

IV. ACCESS

What Finding Aids are Provided?

Information about the holdings are available in multiple levels of detail and by multiple sources as a way to provide various audiences with information about the CenterÕs holdings. The least specific detail is available in the 1996 three volume *Guide to Federal Records in the National Archives of the United States* where electronic records series are described in the context of the larger holdings from a creating organization. Other collective descriptions include *Information About Electronic Records in the National Archives for Perspective Researchers*, General Information Leaflet 37, which also is available on the CenterÕs homepage (http://www.nara.gov/nara/electronic), and a title list of data sets available on the CenterÕs homepage and as a printout. Specific electronic records series descriptions were created as formatted metadata for a portion of the CenterÕs holdings for inclusion in a proposed automated description data base which has not been implemented. The most detailed description for any data set is the documentation package. Each documentation package may contain a narrative describing the data file(s), the record layout and codes for the data, a methodology, sample input forms and questionnaires, annotations regarding the data validity, and a bibliography. The Center also has established an email site (cer@nara.gov) for queries regarding the CenterÕs holdings and services.

Security.

All of the CenterÕs holdings are maintained in environmentally controlled closed stacks which are accessible only by Center staff. Master and backup copies of the data are stored in separate vaults in separate locations to facilitate disaster recovery. The CenterÕs national security classified data sets are in separate environmentally controlled stacks

approved for the storage of classified information. All processing is performed in limited access processing rooms at NARA or at the National Institutes of Health computer center. Computer processing is done on closed systems which require both a registered logon and personal identification number or password to access the system. Researchers do not have direct access to any accessioned data. Presently they access copies of the data that they have purchased for their own use.

Customer Service/Support.

The Center has a staff dedicated to providing reference services to the public and to the staffs of other federal agencies. The staff responds to both general and specific inquiries by telephone, letter, email, or in-person visit and fills orders for copies of specific data and their documentation.. The staff also provides information from records to respond to researcher requests such as for casualty records from the Korean and Vietnam conflicts. The staff also functions as a filter between researchers and the agencies that created the data when problems develop in understanding or interpreting the data. The staff develop a variety of informational material about the CenterÕs holdings and services, much of which is available online.

V. DISSEMINATION

Do You Support Subscriptions?

The Center will accept a standing order (subscription) for electronic records that it receives on a regular, periodic basis from agencies of the Federal government. Under current NARA regulations all subscriptions must be prepaid prior to shipment of the data.

What Media/formats do you use?

Currently the Center provides copies of data files on either nine-track open-reel magnetic tape or class 3480 magnetic tape cartridges encoded in ASCII or EBCDIC, labeled or unlabeled and written to the maximum block size requested. The Center also can provide an exact copy of records in nonstandard formats, if the agency transferred them this way, but it cannot validate or verify the contents of these files. In the past these other formats include packed decimal, zone-decimal, binary, National Information Processing System (NIPS), Statistical Analysis Software (SAS), Statistical Package for the Social Sciences (SPSS), or OSIRSIS. Within the next fiscal year the Center will expand its media options to include diskettes for smaller data sets and CD-ROM. On-line transfer of data remains a more distant goal.

What Transformation (Value Added) is Provided?

The Center currently preserves data as received from the creating agencies; it does not

routinely provide extracts from the data or other value-added services beyond computer validation of the data contents and enhanced documentation. Planned enhancements provide for value-added services including extracts from the data.

Pricing Policies.

The Center uses a cost-recovery fee schedule developed by the National Archives Trust Fund. Currently, the charge for an exact copy of all data on a input cartridge or reel, regardless of the number of data sets on the media is $80.75 when copied to a class 3480 magnetic tape cartridge and $90.00 when copied to a nine-track open reel magnetic tape. The Center charges an additional $24.50 for each additional data set or file added to a media and $7.50 for each subsequent magnetic tape cartridge and $17.00 for each subsequent open-reel magnetic tape. Paper reproductions cost $0.25 per page.

Security.

The same security considerations developed in relation to Access apply to Dissemination. The CenterÕs national security classified data is made available only to researchers who have both the appropriate security clearances and the appropriate need-to-know. Other restricted data are made available only with prior written approval of the creating agency or under the terms of the restrictions which must be supported as a legitimate exemption under the Freedom of Information Act.

VI. SPECIAL CHARACTERISTICS

NARAÕs Center for Electronic Records has a diverse collection which reflects the diverse activities of the federal government. The staff shape the holdings through the process of scheduling, appraisal and accessioning. Currently, the Center acquires less than one percent of all federal records created in an electronic format. The timing of the transfer of electronic records from the creating federal agency to NARA is negotiated with the creator to ensure that the records are available for agency use for as long as necessary for current business and that they are transferred to NARA as soon as practicable to ensure their long term preservation for secondary use. NARA is the only federal agency with an explicit archival mandate for Federal records and thus the only Federal agency that preserves and provides access to a wide range of historically valuable records for the indefinite future. As such it is an archives of last resort for the electronic records of some federal agencies which undertake an active data dissemination function while there is a researcher interest in the data but whose mandate ceases or may cease once the demand wanes or ceases.

A.3 LIFE SCIENCES DATA ARCHIVE

I. What is the domain and who are the customers of the Archive and who are the producers of the data?

The Life Sciences Data Archive (LSDA) Project is responsible for collecting and disseminating data of NASA funded Life Sciences space flight investigations. The LSDAÕs primary customer is the Life Sciences research community, but it is also used by students, educators and the general public. The data archived in the LSDA is produced by

both intramural and extramural investigators funded to perform flight experiments through NASA grants. It is anticipated that the archive may grow to include data from intramural and extramural investigations which are completely ground based.

The LSDA is a distributed archive with responsibilities distributed to LSDA Nodes at various NASA Centers and Projects with Life Sciences activities. There are the LSDA Project Nodes which are responsible for the actual collection and cataloging of data, and there is a LSDA Data Distribution Node which is responsible for dissemination of the data to the public.

II. Ingest Process and Ingest Interface

There are two major types of data producers, first there are the NASA Flight Project offices that design hardware and implement the experiment, and secondly, there are the NASA funded Principal Investigator (PI).

To acquire data from the first type of data producer, the NASA Flight Project offices, the LSDA Project Nodes work closely with them to acquire the data during the flight. The LSDA assists the NASA Flight Project Offices in distributing this data to the PIÕs and gathering it as an archival product.

To acquire data from the second type of data producer, the Principal Investigator, there are a couple of methods of data collection currently being used depending on the ÒageÓ of the experiment. For previously flown experiments (flown prior to 1994) there is a submission agreement between the LSDA and the PIÕs that is based on cooperation, and is not binding. For experiments being selected for flight (after 1994) the funding agreements include a contractual stipulation that the PI must supply the LSDA with raw data, analysed data and a final science report.

These funding agreements are reached when proposed investigations are selected for flight.

At this time the PIs are sent a letter that informs them, that upon the acceptance of funding they will be responsible for delivering the data collected as part of their investigation in a form usable by the sciences community one year post flight.

After a yearÕs proprietary period, the submission of data to the LSDA begins. To assist in its submission, the LSDA Project nodes send the PI an inventory sheet listing expected data. The PI fills out the inventory sheet and returns it to the LSDA Project Node with information as to when the data submissions might occur. In order to clarify the Òusable formÓ requirement throughout the entire LSDA project, the LSDA is in the process of developing a post flight data reporting handbook that will explain exactly how the data should be provided to the archive.

The information objects collected and archived by the LSDA consist of measurements of various parameters in spreadsheet form, electron-micrographs, echocardiographs, video tapes, RACAL tapes, analog tapes, physical specimens, microscope slides, photographs, and hard copy logbooks, lab books and other documents.

The data collections are compiled per investigation, i.e., all data elements, i.e., Information Objects, for a single experiment make a collection. During a typical ingest session each collection is cataloged and catalog information, i.e., metadata, is produced for all the data elements received. This catalog information is entered into a database that is comprised of LSDA approved fields and use of valid values where available. The catalog information is developed by the LSDA personnel at the LSDA Project Node responsible for obtaining the data. The catalog information provides layers of metadata for the data collection that describe the experiment, mission, hardware, personnel, sessions, biospecimen, and research subjects from which the data element was collected. Catalog information is also created for the data element that describes the data element, the type of media, its location, its availability etc.

After the LSDA Project Node creates this catalog information or metadata set for the data collection and the individual data elements, the information goes through a validation process. This post-entry validation is accomplished by a second check of the data by the

LSDA Project Node Manager. Content validation is further ensured by sending the completed catalog entries to the data originator (PI, Flight Project Offices) for verification.

The LSDA has strict security concerns for data from human subjects which require sensitivity and secure handling due to the Human Data Privacy Act. Human data when received is coded to protect the identity of the crewmembers it was collected from. Security procedures include keeping the data on magneto-optical disks stored in a locked file cabinet in a cipher locked room. Overall security procedures stipulate that all digital data are backed up on a daily basis with off-site storage. Access to on-line servers is controlled through the use of password and/or address port filtering.

In most cases a set of data is kept in its original submission form. There are exceptions when the data submitted is on outdated media, for instance, and it is transferred to current media.

III. Internal Forms

The LSDA back-up and storage procedures vary between LSDA Node types. Currently, the LSDA Data Distribution Node resides at the NSSDC. At the NSSDC the LSDA Master Catalog and on-line data reside on a DEC Alpha on a hard drive. These are backed up to tape daily via NSSDCÕs Standard Operating Procedures. At the LSDA Project Nodes most of LSDAÕs data and metadata are stored on magnetic disks and backed up to tape. Long term storage is provided on CD-ROM.

The LSDA is still in a developmental phase and data migration is ongoing. Therefore as yet we have no official migration policy. However, the LSDA Project Nodes are in the process of converting information on outdated media (RA60Õs, RL02Õs) to CD-ROM format.

IV. Access

Access to LSDA data is handled through the LSDA Data Distribution Node. Users enter the LSDA through the World Wide Web (WWW). They are provided a Master Catalog which is comprised of an relational database with a WWW forms interface. The catalog allows users to search metadata to find data that meet their needs. Users are given various areas in which to search. Most users locate data by searching for a particular experiment and then viewing the information (metadata) provided about the data. Data and documents also can be searched for via hardware, research subject, mission, personnel, sessions and biospecimen.

The LSDA does not have any security concerns for users accessing the Master Catalog and non-human digital data. It is freely available to anyone on the WWW. However, human data does have sensitive information, and therefore, there are security concerns. The human flight experiment data is subject to the Human Data Privacy Act, and therefore, security measures are required to control access to this data. The policies and procedures for access to the data are currently being developed.

The LSDA provides user support for questions and problems concerning the Master Catalog (on-line data request system) and for questions about the data being provided. The primary means of user feedback and support is through the LSDA Data Distribution Node . Questions are addressed to the LSDA through on-line ÒWhat do you think?Ó links located throughout the system. From these links a WWW forms interface allows users to submit questions for support. Specific questions about the data are currently addressed by the NSSDC Life Sciences Acquisition Scientist and the LSDA Program Scientist. Questions which can not be answered by either party are forwarded to the LSDA Project Node which collected the data. In some instances to the PI or NASA Flight Project Office who provided the data is contacted to answer the question.

V. Dissemination

The LSDA does not support subscriptions since the system is freely accessible by all users. The LSDA data is located using a catalog on the WWW. Most data is disseminated to the

user through links in the catalog to an anonymous FTP site from which the data is downloaded. This means of data dissemination is, therefore, tightly linked to the data ÒfindingÓ process. If data are non-digital format but are reproducible (i.e., hardcopy documents, or log books) users may request them through on-line ordering forms available in the Master Catalog. The requested information is reproduced via photocopying and shipped via the US Mail to the requester.

The LSDA does contain unique non-reproducible pieces of data such as microscope slides and space flight biospecimens. These unique resources are provided to a requester after a scientific proposal has successfully undergone peer review. Microscope slides may be borrowed from the archive and returned. Biospecimens once disseminated are used to produce original data which is then ingested into the archive.

Currently, the LSDA does not provide many value added services. The data is stored and disseminated as provided by the data producer. Data are available in raw and summarized form. These summarized data are provided by the data producer. LSDA does convert data which are received in a non-standard format are to a more usable form. Currently there are no data analysis tools available through the LSDA. However, the LSDA Project Nodes do ensure that all data sets have the minimum amount of information needed for understanding. (e.g. explanation of all column headings are provided for the spreadsheets, etc.). These are the only Òvalue addedÓ processes that come from the raw data.

Since the LSDA data is primarily disseminated via the WWW, all data are on-line and free. There currently is no charge for data or documents. However, if significant requests are generated for hardcopy documents a processing fee for copying the document may be charged. As yet this has not been determined. In the future CD-ROMs with data may be generated. These CDs will be priced in order to recoup production and distribution costs.

Since the Master Catalog and non-human data in the LSDA are available to anyone on the WWW there is no special security in place for the dissemination of data. However human

data does have sensitive information, and therefore there are security concerns. Limited dissemination of human data will be allowed using the policies and procedures that are currently being developed.

VI. Special Characteristics of your archive

LSDA contains animal, plant, and human space flight data. LSDA is a unique archive in the sense that it provides both digital and non-digital information. This data can be both reproducible and non-reproducible. The non-reproducible data are therefore one of a kind, unique data that can not be duplicated.

Annex B: Archival Information Migration Issues

FROM THE PERSPECTIVE OF THE OAIS, EVEN THE MOST COMMON PHYSICAL AND LOGICAL REPRESENTATIONS OF ARCHIVED INFORMATION EVOLVE AND ARE EVENTUALLY REPLACED. AN OAIS ARCHIVE MUST TO BE READY TO DEAL WITH THE IMPLICATIONS OF THIS EVOLUTION. MOST IMPORTANTLY, THE OAIS MUST UNDERSTAND ALL OF THE REPRESENTATIONS (PHYSICAL AND LOGICAL) USED TO ACQUIRE, STORE, AND DISSEMINATE THE THE DIGITAL INFORMATION FOR WHICH IT IS RESPONSIBLE. IT NEEDS TO BE COGNIZANT OF THE EXTENT TO WHICH EACH OF THOSE REPRESENTATIONS CONTINUES TO HAVE COST-EFFECTIVE HARDWARE AND/OR SOFTWARE SUPPORT AND THE POSITION OF THE REPRESENTATIONS WITHIN THE PRODUCER AND CONSUMER COMMUNITIES. WITHOUT THIS UNDERSTANDING, THE ARCHIVE CAN NEITHER PRESERVE OR PROPAGATE ITS INFORMATION CONTENT.

THERE ARE TWO APPROACHES FOR DEALING WITH THIS EVOLUTION. ONE APPROACH IS TO ACCEPT THAT A REPRESENTATION WILL NOT BE WELL SUPPORTED BY FUTURE COMMON SYSTEMS AND TO ELECT ONLY TO ENSURE THAT THE REPRESENTATION IS WELL DOCUMENTED AND THAT THIS DOCUMENTATION REMAINS FULLY ACCESSIBLE. THE SECOND APPROACH IS MIGRATION AND IT IS THE ISSUES OF THIS SECOND APPROACH WHICH ARE ADDRESSED HERE.

B.1 POLICY AND STORAGE REPRESENTATIONS

When an archive receives a piece of media containing an information set from an producer, there are differing approaches that can be taken to managing the representations and to making the information available to the consumers. Ignoring the case where the archive only maintains the original physical media, the following two cases outline the two extremes on the continuum of approaches.

For the first case, the archive elects to preserve all the representations above the bit representation. In ensures that each representation is either widely understood by the commonly available hardware and software, or that it is documented to the extent needed by the customer community.

For the second case, the archive establishes its own internal set of standard representations down to the bit level, and this includes the inter-relationships of the representations. These representations and their relationships are intended to be able to capture, represent, and retain all the critical information. The critical information in the information set is that information which it is deemed essential to preserve. Upon receipt of

a media unit containing an information set, the archive would move the critical information from this transport media on to an archive media stored in its standard internal representations. The archive would then be able to make subsets of this information available to customers using its internal representation, or it may provide a conversion to a customer requested format before delivery.

For example, consider the case where a number of text files will be delivered to the archive using the EBCDIC mapping for bits and the archive has decided that the essential information is the concept of the letters, not the particular bit stream. The archive would then store the letters in their internal representation, say UNICODE. A customer who only had access to hardware and software which could interpret the ASCII representation may then ask the archive to deliver the text files in this converted form.

Most archives will fall between these two extremes. The important point is that archives need to know which information contained in an information set is essential to preserve and to adopt ingest, migration, and dissemination strategies which will preserve this essential information.

B.2 MIGRATION STRATEGIES WITHIN AN OAIS

All migrations will involve, at a minimum, splitting a particular representation's concepts and inter-relationships from its current mapping to underlying data forms, and mapping them to new underlying data forms. The result is a new representation of the same concepts and inter-relationships. Using language as an example, this is equivalent to taking the concepts expressed in one language and re-expressing them in another language.

There are two major classes of information migration within an archive. The first is the physical preservation of the digital data bit-stream by copying the bit-stream to a new physical media which may or may not be of the same physical type. The second class of information migration involves some sort of change to the data, its organization, or its associated meta-data. We have labeled these two classes "Bit-for-bit" and "Information Unit" Migration.

[SAV: suggestions for better labels.]

B.2.1 BIT-FOR-BIT MIGRATION

[SAV: I need to better understand the IEEE Reference Model for Open Storage Systems Interconnection (IEEE Mass Storage Reference Model) before writing this section]

B.2.2 INFORMATION UNIT

Using the interfaces defined, there are more then one route which may be used to carry out a migration. The method chosen will depend on factors such as the amount of data and the impact on the operational status of the archive.

B.2.2.1 UPDATING AN INTERNAL FORMAT

Consider the case where an archive wishes to change a collection of text files from EBCDIC based content to ASCII based content and where the purpose is to "retire" the EBCDIC dataset on a short timescale. B.2.2.1.1 Consumer/Producer Interface

In this case, the archive policy board might decide that it was more efficient to re-ingest the data in the new format as its own independent collection.

In such a case, a conversion group external to the archive itself could be assigned the job to producing the new data collection. This group would play the role of both consumer and producer within the reference model. As a consumer, the group would request a copy of the collection, including any relevant meta-data and documentation. This now external copy of the collection would be converted from its EBCDIC coding to an ASCII coding. A description of the conversion process and the updated copy of the meta-data and any other documentation associated with the data collection would be updated to reflect the new ASCII coding of the collection. As a producer, the group would provide the new collection to the archive for ingest.

Once the new collection was ingested, archive management could decide whether or not to retire the EBCDIC based collection; or perhaps simply move it to a less accessible part of archive for historical tracking.

B.2.2.1.2 ARCHIVE MANAGEMENT

[SAV: This case will definitely have to be re-written to be more specific once the lower level interfaces are defined.]

In the case of a large collection where it could take a significant amount of time to do the conversion or where through archive policy, it has been decided that the both the EBCDIC and ASCII versions should co-exist at least for the foreseeable future, moving the entire collection outside of the archive may not be possible. In such a case, the conversion would have to be done within the confines of the archive.

The Archive Management Entity is the only one in the reference model which has access to all of the parts of the archive needed to perform such a conversion. It is assumed that the process described is part of Archive Management.

It is also assumed that it has been decided that both the EBCDIC and ASCII versions of the data should appear as part of the same collection. The first step then is to make modification to the meta-data holdings to indicate that a ASCII copy may be available and to update any Access and Dissemination Entities that need to be taught about the new data format. Once this new setup has been completed, a process within Archive Management can be defined which would operate as follows. Based on priorities, the process would be sent a list of elements within the collection which needed to be converted. The process would retrieve the related meta-data and request that the Storage Entity transfer the specific data element to closed Storage area. The EBCDIC to ASCII conversion would then take place at this physical location under control of the Archive Management Entity. After this was complete, the Storage Entity would be ordered to permanently store the converted data and delete the temporary copies. Upon receiving this confirmation, the Archive Management Entity would send the appropriate updates to the Database Management Entity. This process would continue until the list of data elements to convert was complete.

Such a scenario also has the advantage that it could be made a standing process within Archive Management. Consider the case where due to resource limitations or their own internal policy a producer is only able to deliver a collection in EBCDIC. As an archive policy, it could be decide that the data collection should also be stored in ASCII. This policy could either be implemented by having the Ingest Entity convert the data before initial loading or the above method could be used once the Archive Management Entity was taught to trigger upon notice of a new ingest of EBCDIC data.

[SAV: An option we might wish to consider is defining some sort of on-the-fly conversion capability within Dissemination. If we did, another scenario would be to have only "management" functions for the conversion in Archive Management and use the Dissemination on-the-fly for also converting data elements for internal

updates/conversions. This would also eliminate the possible conflict above of having a Archive Management process use physical space within Storage for an active operation.]

ANNEX C: COMPATABILITY WITH OTHER STANDARDS

Because of the similarities between the access and dissemination functions of digital archives and digital libraries, this section extends the concepts presented in the Z39.50 Profile for Access to Digital Collections and the companion profile, Z39.50 Profile for Access to Digital Library Objects, to include the features needed to fulfill the requirements of the long-term information preservation function discussed in Section 2 of this paper.

The Z39.50 Profile for Access to Digital Collections is currently under development for the U.S. Library of Congress. The intent of the profile is to provide a very high-level structure to enable a user to navigate thematically organized, hierarchically structured collections of descriptions of digital and physical objects. Work on the specification of this profile was begun on in September 199 5, and draft version 7 of the digital collections profile (DCP) was released on May 3,1996. Because of the focus on high-level compatibility, the DCP has several stated limitations, which are intended to be addressed in companion profiles that deal with more specific domains.

Figures C-1 and C-2 are OMT diagrams derived from Sections 2 and 4 of the DCP. The DCP combines both a logical view of digital collections and a physical view of data collections. Figure C-1 is derived from the DCP logical view, in which collections are composed of objects and subcollections and reference other collections of interest (related collections).

- The DCP treats digital objects as atomic, that is, their content is opaque. Thus the profile addresses searching descriptive information rather than searching digital objects.

- The DCP treats descriptive items (e.g., finding aids, cataloging records, and exhibition catalogs) as opaque, though clients may have at their disposal helper applications that are able to process or display them.

¥ The DCP does not model complex relationships among objects of all classes.

The DCP of the holdings of an archive is based on the basic concepts of description information being logically separated from the described digital object, the description records or item descriptors being organized into collections, and the members of a collection being objects or other subcollections. The link to the physical model shown in Figure C-2 is the concept of Descriptive Records (both for Objects and Collections) and the concept of a Datastore, which comprises the set of all the portions of databases that make up a collection.

Figure C-2. DCP: Collection Viewpoint

The central concept of the physical view of the DCP shown is the Descriptive Record. A Descriptive Record exists for a collection, and each of the collections and objects are contained within the collection. While a collection is represented as a Datastore composed of multiple databases that may be on different servers, a Descriptive Record must be contained in a single database and server. A Collection Descriptive Record may enumerate all contained objects and collections. The schema for a Descriptive Record is stated in Section 4 of the DCP. A Descriptive Record contains all the associate descriptions for the object or collection it describes and either a pointer to the object(or collection) or the digital object itself. The Descriptive Record can be considered as a database record, a retrieval record, or an abstract database record (schema), depending on the context in which it appears. The DCP states that the distinction between whether something is an object or associated description is dependent on the viewpoint of the collection producer.

Figure C-3. DCP Physical View

The Digital Library Profile (DLP) is a companion profile that extends the DCP to include features needed for access to digital libraries. The primary augmentation to the DCP is the fact that the digital object class is no longer opaque. The digital objects in the DLP are trees in which each nonleaf node has an arbitrary number of subtrees and/or leaves, and leaf nodes represent data. Each node has a string tag whose purpose is to convey to the user what that node represents. If the string is not sufficiently descriptive, a descriptive meta-element may also accompany any node.

The OAIS extends the DCP by replacing the ÒobjectÓ class shown in Figure C-1 with the Òarchive information packageÓ shown in Figure 3-a. In addition to the Òdata object,Ó which is the object (digital or physical) that the archive is tasked to preserve, the Òarchive information packageÓ contains a significant amount of metadata and ancillary data required to preserve the meaning of the Òdata object.Ó This information can be divided into the following categories, which reflect the categories discussed in the paper ÒPreserving Digital Information:Ó

Annex D: Brief Guide to the OMT

(TBD)